目　录

第一章 插 值

§1.1 引 言

在生产实践的许多领域里，例如机械工业、造船、汽车制造，常常有这样的问题：给了一批离散样点，要求作出一条光滑曲线(乃至于曲面)，使其通过或尽可能地靠近这些样点，以便满足设计要求或者据此进行机械加工。在过去，这种放样工作大都是用人工方式进行的。为了提高劳动生产率、提高设计质量，日益需要在计算机的协助下，用数学的方法自动进行。这就是所谓数学放样以及曲线、曲面的自动产生的问题。

另外，在使用计算机解题时，由于机器只能执行算术的和逻辑的操作，因此，任何涉及连续变量的计算问题都需要经过离散化以后才能解算。例如计算积分时，采用离散点函数值累加，即数值积分的方法，计算微分用差商即数值微分的方法，解微分方程用格网即差分方法，以及有限元方法等等，也都直接或间接地要用到在离散数据的基础上补插出连续函数的思想和方法。

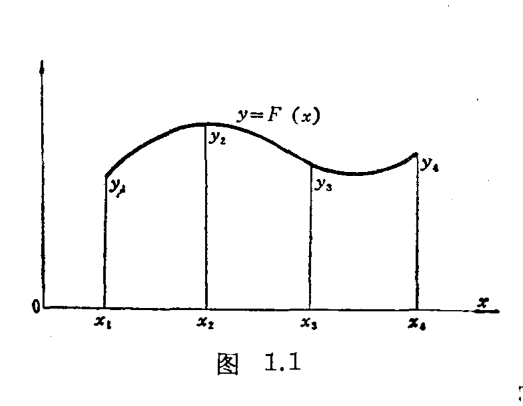


图 1.1

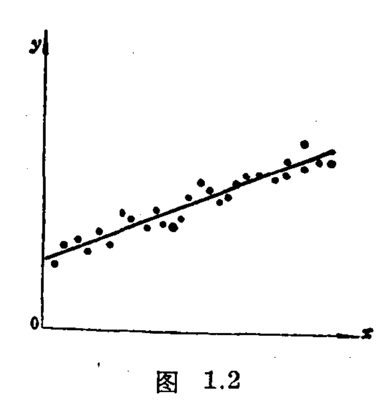


图 1.2

与此相关的一类数学问题是插值问题，即当原始数据(x_0, y_0)，(x_1, y_1)，$\cdots$，(x_n, y_n)是精确的或者可靠度较高时，要求定出一个便于计算的"初等"的函数或曲线$y=F(x)$(例如多项式，或者分段多项式等等)通过给定的离散样点

$$F(x_0)=y_0, \quad F(x_1)=y_1, \cdots, \quad F(x_n)=y_n$$

如图1.1。这时待定函数的自由度，即待定参数(例如多项式的系数)的个数与给定的插值条件个数相当。这就是本章的主题。

另一类是最优拟合或最小偏差问题，即当原始数据本身含有"噪音"，即含有不可避免的误差时，要求定出一个初等的函数$F(x)$，不是要求严格地通过样点，而是要求最优地靠近样点，即在某种意义下总的偏差为最小(图1.2)。这时待定函数的自由度恒小于、甚至于远小于样点个数，从而可以达到滤去噪音的目的。这将在§1.6和第四章中讨论。

§1.2 多项式插值

为了讨论的方便，统一约定一些记号和名词如下：恒设有某个原函数$f(x)$，它在一些节

点 $x=x_0$, x_1, …处的值为 $f_0=f(x_0)$, $f_1=f(x_1)$, …。必要时还引用导数值 $f'_0=f'(x_0)$, $f'_1=f'(x_1)$, …。命 $F(x)$ 为适当的插值多项式。原函数与插值函数的差 $f(x)-F(x)$ 也叫做插值余项。符号 $[x_0, x_1, \cdots]$ 表示含有点 x_0, x_1, …的最小区间，也就是以 $\max(x_0, x_1, \cdots)$ 和 $\min(x_0, x_1, \cdots)$ 为端点的区间。

1.2.1 最简单的插值公式

介绍几种简单常用的插值公式。它们可以启示一般插值(1.2.2节)的作法，也是分段插值(§1.3)的基础。为了方便，在各个公式后都附列余项估计，但不加证明。

(一)一点零次——水平插值

过样点 (x_0, f_0) 作水平线(图 1.3)，即

$$F(x)=f_0 \tag{1.2.1}$$

$$f(x)-F(x)=f'(\xi)(x-x_0),\ \xi\in[x_0, x] \tag{1.2.2}$$

这是零次插值，对于函数值在插点邻近有一阶精度，但对于导数则没有逼近性。

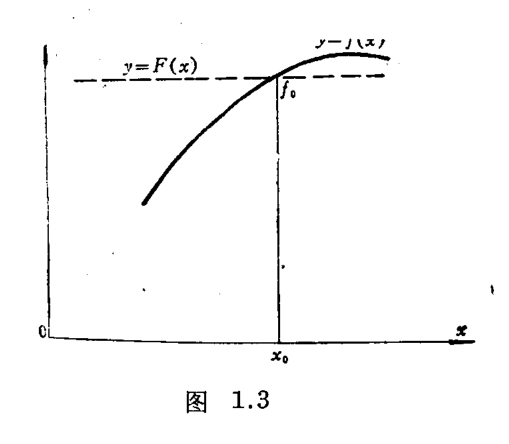


图 1.3

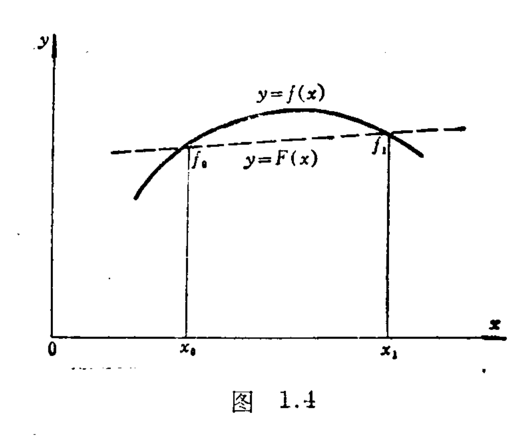


图 1.4

(二)两点一次——线性插值

过两个样点 (x_0, f_0), (x_1, f_1) 作直线(图 1.4)，即

$$F(x)=\frac{x-x_1}{x_0-x_1}f_0+\frac{x-x_0}{x_1-x_0}f_1 \tag{1.2.3}$$

$$f(x)-F(x)=\frac{1}{2}f''(\xi)(x-x_0)(x-x_1),\xi\in[x_0, x_1, x] \tag{1.2.4}$$

这是一次插值，对于函数值有二阶精度，同时对于导数也有了逼近性。

为了分析的方便，可以命

$$l_0(x)=\frac{x-x_1}{x_0-x_1},\quad l_1(x)=\frac{x-x_0}{x_1-x_0} \tag{1.2.5}$$

显然有

$$l_0(x)+l_1(x)\equiv 1$$

$$l_0(x_0)=1,\quad l_0(x_1)=0,\quad l_1(x_0)=0,\quad l_1(x_1)=1$$

而线性插值可以表为

$$F(x)=f_0l_0(x)+f_1l_1(x) \tag{1.2.6}$$

函数 $l_0(x)$, $l_1(x)$ 可以称为线性插值的基函数。

(三)三点二次——抛物插值

过三个样点(x_0, f_0), (x_1, f_1), (x_2, f_2)作抛物线(图1.5),在解析上可以表为

$$F(x)=f_0l_0(x)+f_1l_1(x)+f_2l_2(x) \tag{1.2.7}$$

这里插值基函数是

$$l_0(x)=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)},\quad l_1(x)=\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)},\quad l_2(x)=\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \tag{1.2.8}$$

而插值余项为

$$f(x)-F(x)=\frac{1}{3!}f'''(\xi)(x-x_0)(x-x_1)(x-x_2),\ \xi\in[x_0,\ x_1,\ x_2,\ x] \tag{1.2.9}$$

很容易验证

$$l_0(x_0)=1,\quad l_0(x_1)=0,\quad l_0(x_2)=0$$
$$l_1(x_0)=0,\quad l_1(x_1)=1,\quad l_1(x_2)=0$$
$$l_2(x_0)=0,\quad l_2(x_1)=0,\quad l_2(x_2)=1$$

因此函数(1.2.7)确实满足插值条件

$$F(x_0)=f_0,\ F(x_1)=f_1,\ F(x_2)=f_2$$

从图1.5以及余项估式(1.2.9)可以看出，这里逼近程度又提高了。函数值有三阶精度，并且直到二阶导数都有逼近性。

以上几种都是以节点的函数值f_0, f_1, …为基础的插值，即所谓拉格朗日(Lagrange)插值。当在节点上除了函数值f_0, f_1, …外还掌握导数值f'_0, f'_1, …时，则还可以采用带导数的插值，即所谓埃尔米特(Hermite)插值。除了"过点"外还要求"相切"，密合程度就会更好些。举出两种最简单的情况:

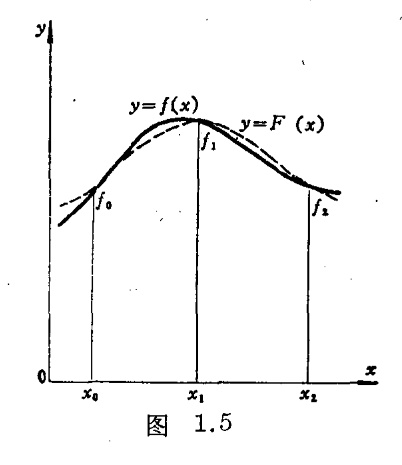


图 1.5

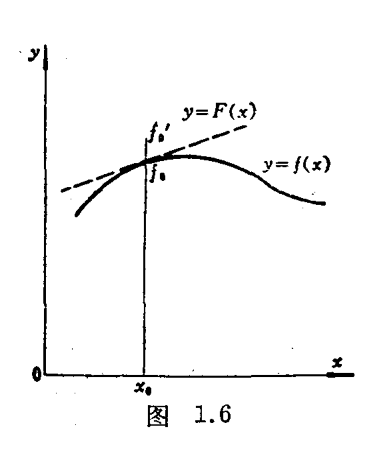


图 1.6

(四)一点一次带导数插值

$$F(x)=f_0+(x-x_0)f'_0 \tag{1.2.10}$$

$$f(x)-F(x)=\frac{1}{2}f''(\xi)(x-x_0)^2,\ \xi\in[x_0,\ x]$$

这就是切线插值(图1.6)。显然式(1.2.10)满足插值条件

$$F(x_0)=f_0,\quad F'(x_0)=f'_0$$

(五)两点三次带导数插值

要求作三次多项式

$$F(x)=a_0+a_1x+a_2x^2+a_3x^3 \tag{1.2.11}$$

满足(图1.7)

$$F(x_0)=f_0,\quad F'(x_0)=f'_0,\quad F(x_1)=f_1,\quad F'(x_1)=f'_1 \tag{1.2.12}$$

利用这四个条件可以解出四个系数 a_0, a_1, a_2, a_3, 它们都线性地依赖于 f_0, f'_0, f_1, f'_1。这里演算是初等的，但比较繁琐，故从略，而直接给出最终表达式

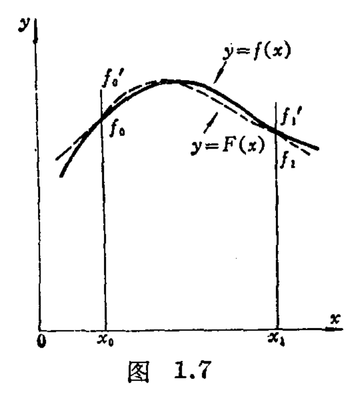


图 1.7

$$F(x)=f_0\alpha_0(x)+f'_0\beta_0(x)+f_1\alpha_1(x)+f'_1\beta_1(x) \tag{1.2.13}$$

这里插值函数 α_i, β_i 可以通过两点线性插值的基函数(1.2.5)

$$\left.\begin{array}{l} l_0(x)=\dfrac{x-x_1}{x_0-x_1},\quad l_1(x)=\dfrac{x-x_0}{x_1-x_0}, \\ l_0(x)+l_1(x)\equiv 1 \end{array}\right\} \tag{1.2.14}$$

表达如下：

$$\left.\begin{array}{l} \alpha_0(x)=l_0^2(1+2l_1)=3l_0^2-2l_0^3 \\ \alpha_1(x)=l_1^2(1+2l_0)=3l_1^2-2l_1^3 \\ \beta_0=hl_0^2l_1=h(l_0^2-l_0^3) \\ \beta_1=-hl_1^2l_0=-h(l_1^2-l_1^3) \\ h=x_1-x_0 \end{array}\right\} \tag{1.2.15}$$

由于(1.2.14)以及

$$l'_0(x_0)=l'_0(x_1)=-\frac{1}{h},\quad l'_1(x_0)=l'_1(x_1)=\frac{1}{h}$$

不难看出

$$\begin{array}{llll} \alpha_0(x_0)=1, & \alpha'_0(x_0)=0, & \alpha_0(x_1)=0, & \alpha'_0(x_1)=0 \\ \beta_0(x_0)=0, & \beta'_0(x_0)=1, & \beta_0(x_1)=0, & \beta'_0(x_1)=0 \\ \alpha_1(x_0)=0, & \alpha'_1(x_0)=0, & \alpha_1(x_1)=1, & \alpha'_1(x_1)=0 \\ \beta_1(x_0)=0, & \beta'_1(x_0)=0, & \beta_1(x_1)=0, & \beta'_1(x_1)=1 \end{array}$$

因此(1.2.13)确实满足插值条件(1.2.12)，它就是所要求的三次插值函数，而余项估计则为

$$f(x)-F(x)=\frac{1}{4!}f^{(4)}(\xi)(x-x_0)^2(x-x_1)^2,\ \xi\in[x_0,\ x_1,\ x] \tag{1.2.16}$$

比以前几种又提高了精度，具有直到三阶导数的逼近性。

表达式(1.2.15)还可以换个写法，命

$$\omega(x)=(x-x_0)(x-x_1)$$

于是

$$\omega'(x_0)=(x_0-x_1),\quad \omega'(x_1)=(x_1-x_0),\quad \omega''(x_0)=\omega''(x_1)=2$$

$$\left.\begin{array}{l} \alpha_i(x)=(l_i(x))^2\left[1-\dfrac{\omega''(x_i)}{\omega'(x_i)}(x-x_i)\right] \\ \beta_i(x)=(l_i(x))^2(x-x_i),\quad i=0,\ 1 \end{array}\right\} \tag{1.2.17}$$

这种形式便于推广到高次埃尔米特插值(1.2.2节)，而形式(1.2.15)则便于推广到分段埃尔米特插值(1.3.1节)。

1.2.2 一般的多项式插值和误差

首先讨论拉格朗日插值。对于 $n+1$ 个节点 $x_0, \cdots, x_n$ 的函数值 $f_0, \cdots, f_n$ 要求作 n 次多项式

$$F(x)=a_0+a_1x+\cdots+a_nx^n \tag{1.2.18}$$

使得

$$F(x_i)=f_i, \quad i=0, 1, \cdots, n \tag{1.2.19}$$

根据 1.2.1 节中式(1.2.3)的启发，可以作 $n+1$ 个基函数

$$l_i(x)=\prod_{\substack{k=0\\k\neq i}}^{n}\frac{x-x_k}{x_i-x_k}, \quad i=0, 1, \cdots, n \tag{1.2.20}$$

显然可见

$$l_i(x_j)=\delta_{ij}=\begin{cases}1, & 当\ i=j\\0, & 当\ i\neq j\end{cases} \tag{1.2.21}$$

因此

$$F(x)=\sum_{i=0}^{n}f_il_i(x) \tag{1.2.22}$$

满足插值条件(1.2.19)，它就是所要求的多项式。也可以命

$$\omega(x)=\prod_{k=0}^{n}(x-x_k) \tag{1.2.23}$$

不难验证

$$\omega'(x_i)=\prod_{\substack{k=0\\k\neq i}}^{n}(x_i-x_k) \tag{1.2.24}$$

因此，基函数(1.2.20)也可以表成

$$l_i(x)=\frac{\omega(x)}{(x-x_i)\omega'(x_i)}, \quad i=0, 1, \cdots, n \tag{1.2.25}$$

利用微积分中的洛尔定理可以证明(见[5])：n 次拉氏插值的余项可以表为

$$f(x)-F(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi)\omega(x), \ \xi\in[x_0, \cdots, x_n, x] \tag{1.2.26}$$

1.2.1 节中式 (1.2.3) 的余项都是此式的特例。命 h 为区间 $[x_0, \cdots, x_n]$ 的宽度，M_{n+1} 为 $|f^{(n+1)}|$ 在这个区间上的极大值，则还可证明，函数连同导数有如下的估计式

$$|f^{(p)}(x)-F^{(p)}(x)|\leqslant\frac{1}{(n+1-p)!}M_{n+1}h^{n+1-p}, \ 0\leqslant p\leqslant n, \ x\in[x_0, \cdots, x_n] \tag{1.2.27}$$

据此，对于(1.2.3)各式的导数逼近性可以得出相当的结论。

埃尔米特插值也可以推广到一般的情况。对于 $n+1$ 个节点 $x_0, \cdots, x_n$ 上的函数值 $f_0, \cdots, f_n$ 及一阶导数值 $f'_0, \cdots, f'_n$，根据 $2(n+1)$ 个插值条件

$$F(x_i)=f_i, \quad F'(x_i)=f'_i, \quad i=0, 1, \cdots, n \tag{1.2.28}$$

可以唯一地定出 $2n+1$ 次的插值多项式 $F(x)$。在 1.2.1 节中式(1.2.13)、(1.2.17)的启发下，可以取 $2(n+1)$ 个基函数

$$\left.\begin{aligned}\alpha_i(x)&=(l_i(x))^2\left[1-\frac{\omega''(x_i)}{\omega'(x_i)}(x-x_i)\right]\\\beta_i(x)&=(l_i(x))^2(x-x_i), \quad i=0, 1, \cdots, n\end{aligned}\right\} \tag{1.2.29}$$

这里 $l_i(x)$, $\omega(x)$ 由式(1.2.20)，(1.2.23)给出。不难验证

$$\alpha_i(x_j)=\delta_{ij},\quad \alpha_i'(x_i)=0$$
$$\beta_i(x_j)=0,\quad \beta_i'(x_j)=\delta_{ij}$$

因此

$$F(x)=\sum_{i=0}^{n}[f_i\alpha_i(x)+f_i'\beta_i(x)] \tag{1.2.30}$$

确实满足插值条件(1.2.28)，它就是所要求的埃尔米特插值多项式。类似于式(1.2.26)，(1.2.27)有余项估计

$$f(x)-F(x)=\frac{1}{(2n+2)!}f^{(2n+2)}(\xi)(\omega(x))^2,\quad \xi\in[x_0,\cdots,x_n,x] \tag{1.2.31}$$

$$|f^{(p)}(x)-F^{(p)}(x)|\leqslant\frac{1}{(2n+2-p)!}M_{2n+2}h^{2n+2-p},\quad 0\leqslant p\leqslant 2n+1,\ x\in[x_0,\cdots,x_n] \tag{1.2.32}$$

埃尔米特插值还可以推广到包含更高阶导数以及各节点包含的导数阶不均等的情况，这里就不赘述了。

关于插值过程的稳定性

根据余项估计(1.2.26)、(1.2.31)，似乎会认为插值的次数愈高愈好，但并不尽然。实际上，在插值过程中误差有两种来源：一是由原函数$f(x)$被代以插值函数$F(x)$引起的，这就是上面说到的余项，即截断误差；另一是由节点数据f_i本身的误差所引起的。通常由于实验的误差，或者计算过程中的舍入等等，总会带来数据误差，这种误差在插值过程中可能被扩散和放大，这就是插值的稳定性问题。

以拉氏插值为例，设真值f_i被代以含误差的$\bar f_i=f_i-\delta f_i$，命$\bar F(x)$为以$\bar f_0,\cdots,\bar f_n$为基础的插值多项式，于是最终的误差是

$$f(x)-\bar F(x)=[f(x)-F(x)]+[F(x)-\bar F(x)]$$

右端第一项就是余项，根据式(1.2.22)第二项表为

$$F(x)-\bar F(x)=\sum_{i=0}^{n}\delta f_i l_i(x)$$

因此，在节点x_i上的数据误差δf_i是通过该点的插值基函数$l_i(x)$而全面扩散乃至放大的。因此，插值基函数也就是数据误差的"影响函数"。图1.8表示一个$l_i(x)$，它在基本区间$[x_0,\cdots,x_n]$内，即内插时作波动状；在基本区间之外则按距离的n次幂放大。因此当n趋大时，插值过程对于样点的数据误差非常敏感，这就是说高次插值具有数值不稳定性。

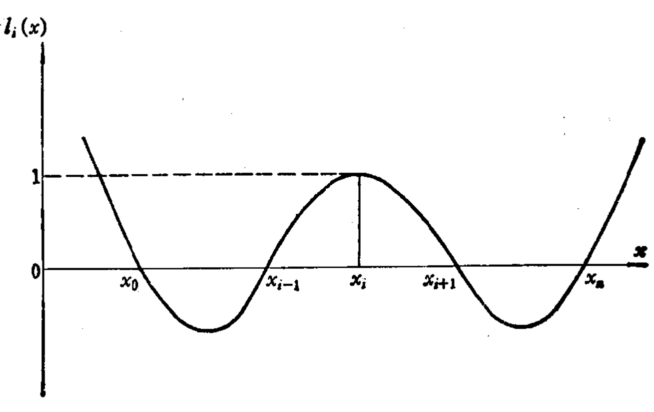


图 1.8

1.2.3 关于多项式插值的运用

单纯的多项式插值是简单而通用的方法，但是它也有其弱点，即具有上面所说的高次时的数值不稳定性。因此，在实际运用时不可盲目使用，否则会导致严重的差错。

(一)提高精度的问题

比较图1.5～图1.9可以看出，当插值的次数提升时，逼近程度也逐步改善。例如二次插值在小范围内就有很好的密合。事实上，当函数$f(x)$在整个实数轴$-\infty<x<\infty$上为解析函数，并且作为复变函数$f(z)$在整个z平面上为解析，则可以证明：在实轴的任意区间$a\leqslant x\leqslant b$上采用等距节点，并且逐次加密的插值过程收敛于原函数$f(x)$，在这种情况下精确度是随次数升高而提高的。

对于不光滑的函数情况就不相同了。例如取$f(x)=|x|$，它在$x=0$处有导数间断，可以证明在区间$|x|\leqslant 1$上逐次加密的等距插值过程是发散的。更有甚者，如取函数

$$f(x)=\frac{1}{1+x^2}$$

(图1.9)，它在整个实轴上解析，因此是高光滑度的函数。但是，可以证明，在区间$|x|\leqslant 5$上逐次加密的等距插值过程是发散的。这一事实与前引结果并不矛盾，因为$f(x)$在复数平面内的延拓$f(z)=\dfrac{1}{1+z^2}$在虚轴上有两个奇点$z=\pm i$，发散性正是由这两个"隐藏"的奇点引起的。图1.10中的实线表示一个通过十个点的九次插值，在区间的两端出现大幅度的波动扭拐，显然是"多余"的，不合理的。

由于上述原因，盲目地提高插值次数是不可取的，甚至会导致极坏的后果。实践中，插值次数高于6、7次就很少了。一般是采用分段低次的插值来提高精度。图1.10中的虚线是用分段三次，即所谓样条插值的结果，比单纯的九次插值有显著改进(参考§1.3)。

(二)外插的问题

当计算点落在插值基本区间之内时叫做内插，否则叫做外插。有时人们仅在变量的一

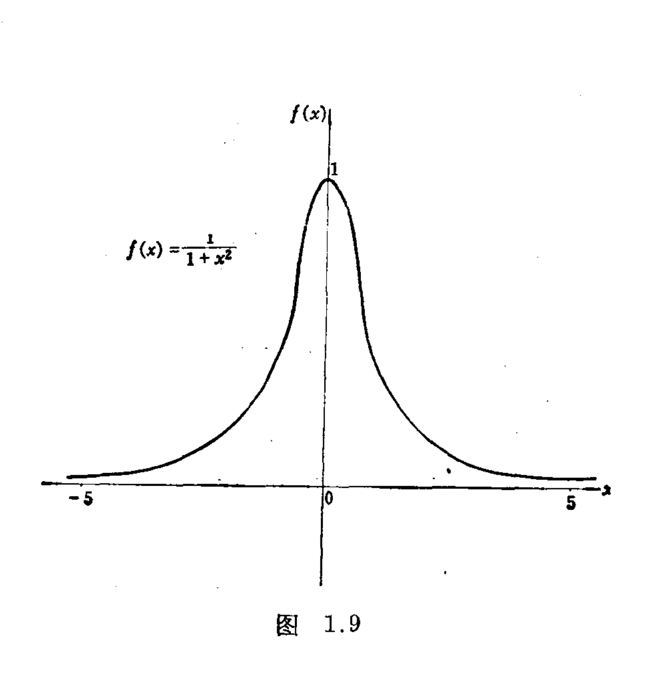


图 1.9

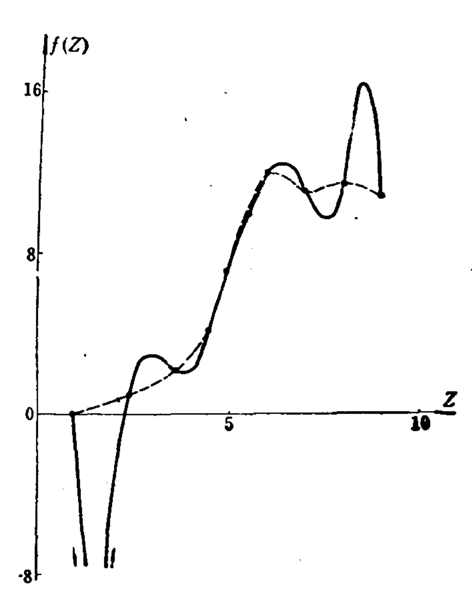


图 1.10

定范围之内掌握数据和规律，需要据此外推在该范围以外的行为，这就要用外插。

据1.2.2节所述，一个节点上的误差的影响，当传出了基本区间之外后，就会按插值次数n的幂次而无穷增长。因此外插，特别当次数较高和距离较远时是很不可靠的。

临边节点的布局也影响到外插的可靠性。例如用三点$x_0=0$, $x_1=\sigma$, $0<\sigma<1$, $x_2=1$外插到$x=1.5$及$x=2$。设在节点x_1处函数值f_1有误差$\delta=\pm1$，则它在各点的影响，即放大因子为$l_1(x)=\mp\dfrac{x(x-1)}{\sigma(1-\sigma)}$。图1.11表示对不同的$\sigma$时的误差曲线。当$\sigma=0.9$或$0.1$时$l_1(1.5)=\mp8.33$, $l_1(2)=\mp22.2$。就好像不等臂的杠杆，在短臂端按下一点点，在长臂端就翘起很多。这种现象在外插是常常出现的(图1.11a)。如取$\sigma=0.5$则情况改善，得$l_1(1.5)=\mp3$(图1.11b)。如果索性抛弃x_2改用x_0, x_1的线性插值，则在x_1处的单位误差影响为$l_1(x)=\dfrac{x}{\sigma}$，仍取$\sigma=0.9$，则情况进一步改善，得$l_1(1.5)=\pm1.7$(图1.11c)。间距的不均匀性，外插点的距离以及插值次数的这些影响，可以从图1.11(a)、(b)、(c)看出。因此进行外插时应该依据对具体问题及插值规律的了解而慎重处理。

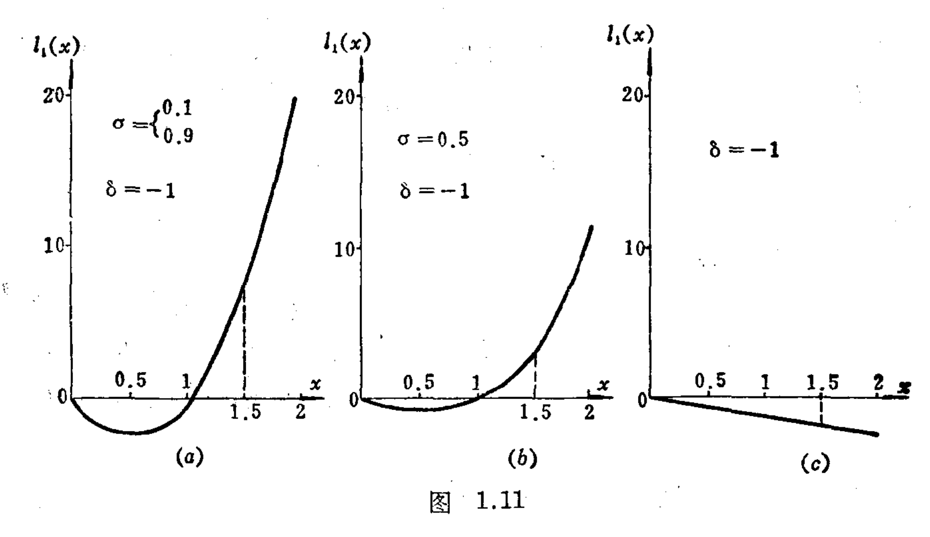


图 1.11

(三)含有间断性的问题

多项式是一种"初等"的解析函数，本身具有最高度的光滑性，因此在本质上只适应于高光滑函数的插值。当原函数自身或其某阶导数有间断时，次数高于间断阶的多项式插值是不适应的。

图1.12中的间断曲线(实线)表示在正态压力下单位质量的水H_2O的热容量C随温度T的变化率。它有两个间断点，即冰点$T=0$℃及沸点$T=100$℃。间断点的跃值分别相应于熔化热及汽化热。曲线的斜率就是比热。如果在固、液、气三相各取一个代表点作二次插值，如图中虚线所示，则显然是相当歪曲了真相的。如果同样取这三个样点，但是分别在三相，那么，即使作0次(即水平)插值，它就大有改进了。

当函数连续而导数间断时，情况也类似。例如：在区间$-1\leqslant x\leqslant1$上，取$f(x)=1-|x|$，用$x=-1, 0, 1$用三个样点作抛物线，如图1.13中的虚线，其情况显然不好。如果逐次等距加密节点作高次插值，则可以证明必以发散告终。但如果仍取$x=-1, 0, 1$为节点，分段作线性插值，则恰好准确地得到原函数。这虽然是个极端的例子，但也足以说明对于插值应该看对象灵活运用，而不要陷于盲目性。

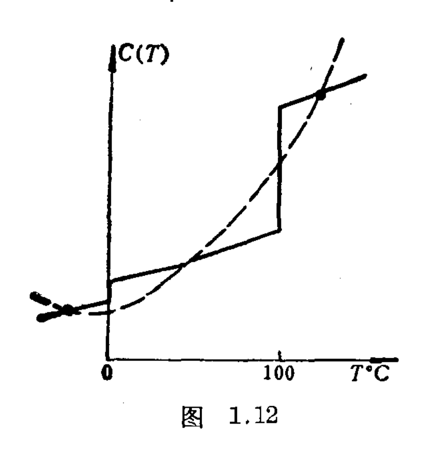


图 1.12

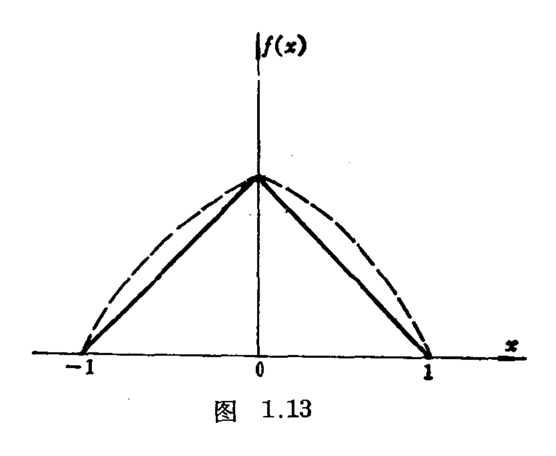


图 1.13

§1.3 片段多项式插值

我们知道，多项式插值在低次时是简单方便的，在高次时则有数值不稳定的缺点。因此，当节点数很多时，自然设想分段用低次插值，而在分点处保证一定的连续性。这类方法通常有较好的收敛性和稳定性，算法也简单。大致可以分为两种类型：一种是局部化的分段插值(1.3.1 节)，这是简单的低次插值公式的直接推广。另一种是非局部化的分段值插，即所谓样条插值(1.3.2 节)，构成稍繁，但光滑度较高。

统一约定在区间$[a, b]$上给了节点

$$a=x_0<x_1<\cdots<x_{n-1}<x_n=b \tag{1.3.1}$$

它们把$[a, b]$剖分为子区间

$$[x_i, x_{i+1}],\ 间距\ h_{i+\frac{1}{2}}=x_{i+1}-x_i,\quad i=0, 1, \cdots, n-1 \tag{1.3.2}$$

设给定了某原函数$f(x)$在各节点x_i的值f_i，必要时还考虑导数值f'_i，要求作分段插值多项式$F(x)$，即$F(x)$在每个子区间$[x_i, x_{i+1}]$上是多项式，但在不同子区间上可以是不同的多项式，而在各连接点x_i保证函数式一定阶导数的连续性。必要时还引用半点(即各子区间的中点)

$$x_{i+\frac{1}{2}}=\frac{1}{2}(x_i+x_{i+1}),\quad i=0, 1, \cdots, n-1 \tag{1.3.3}$$

作为插值节点或作为多项式的分段点。

1.3.1 简单的片段多项式插值

1.2.1 节的几种简单低次插值都可以推广为分段插值如下：

(一) 分段零次——台阶状插值

$$F(x)=f_i,\ 当\ x_{i-\frac{1}{2}}\leqslant x\leqslant x_{i+\frac{1}{2}},\quad i=0, 1, \cdots, n \tag{1.3.4}$$

(约定$x_{-\frac{1}{2}}=x_0$, $x_{n+\frac{1}{2}}=x_n$)。这就是在每个子区间$[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$作水平插值，综合成为台阶状(如图 1.14)，在半点$x_{i+\frac{1}{2}}$有间断。这虽然是比较粗糙的方法，但适合于有间断或有些不规则的情况。这种插值函数也可用基函数来表达，为此，对$i=0, 1, \cdots, n$命

$$\pi_i(x)=\begin{cases}1, & 当\ x_{i-\frac{1}{2}}\leqslant x\leqslant x_{i+\frac{1}{2}}\\ 0, & 它处\end{cases} \tag{1.3.5}$$

如图 1.15，于是

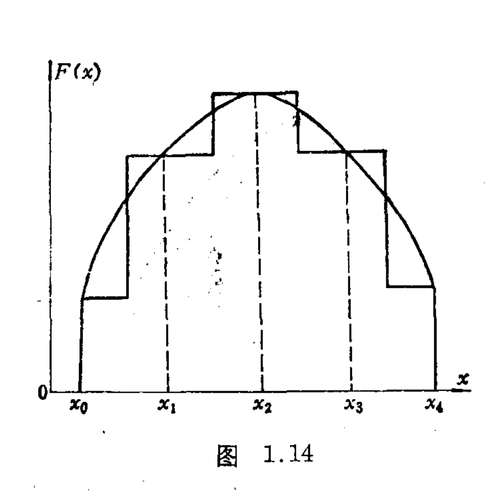


图 1.14

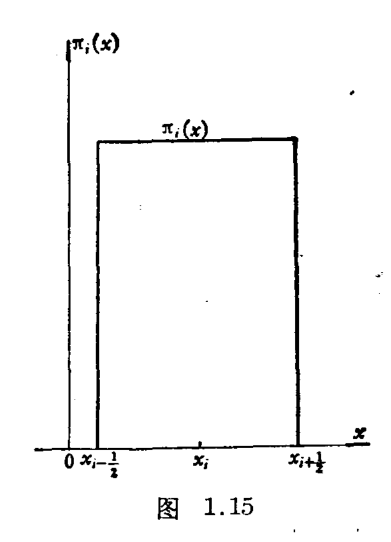


图 1.15

$$F(x)=\sum_{i=0}^{n} f_i\pi_i(x) \tag{1.3.6}$$

也可以把样点值给在半点上：$f_{\frac{1}{2}}$, $f_{1+\frac{1}{2}}$, $\cdots$, $f_{n-\frac{1}{2}}$。于是台阶状插值为

$$F(x)=f_{i+\frac{1}{2}},\quad 当\ x_i\leqslant x\leqslant x_{i+1},\quad i=0,\ 1,\ \cdots,\ n-1 \tag{1.3.4$'$}$$

这时基函数取为

$$\pi_{i+\frac{1}{2}}(x)=\begin{cases}1, & 当\ x_i\leqslant x\leqslant x_{i+1}\\ 0, & 它处\end{cases} \tag{1.3.5$'$}$$

于是

$$F(x)=\sum_{i=0}^{n-1} f_{i+\frac{1}{2}}\ \ \pi_{i+\frac{1}{2}}(x) \tag{1.3.6$'$}$$

(二)分段线性——折线插值

过样点(x_0, f_0), $\cdots$, (x_n, f_n)作折线相连(图 1.16),

$$F(x)=\frac{x-x_{i+1}}{x_i-x_{i+1}}f_i+\frac{x-x_i}{x_{i+1}-x_i}f_{i+1},\quad x_i\leqslant x\leqslant x_{i+1},\quad i=0,\ 1,\ \cdots,\ n-1 \tag{1.3.7}$$

这是分段一次多项式,并在$[a, b]$上连续,但在x_i处导数有间断。它也可以用基函数来表达.为此,对于$i=0, 1, \cdots, n$, 命

$$\lambda_i(x)=\begin{cases}\dfrac{x-x_{i-1}}{x_i-x_{i-1}}, & 当\ x_{i-1}\leqslant x\leqslant x_i\\[2mm] \dfrac{x-x_{i+1}}{x_i-x_{i+1}}, & 当\ x_i\leqslant x\leqslant x_{i+1}\\[2mm] 0, & 它处\end{cases} \tag{1.3.8}$$

(约定$x_{-1}=x_0$, $x_{n+1}=x_n$)。显然$\lambda_i(x)$是分段一次的连续函数并满足

$$\lambda_i(x_j)=\delta_{ij} \tag{1.3.9}$$

如图 1.17 所示,于是

$$F(x)=\sum_{i=0}^{n} f_i\lambda_i(x),\quad a\leqslant x\leqslant b \tag{1.3.10}$$

(三)分段二次插值

在每个子区间$[x_i, x_{i+1}]$上作三点二次插值，即过三个样点(x_i, f_i), $(x_{i+\frac{1}{2}}, f_{i+\frac{1}{2}})$,

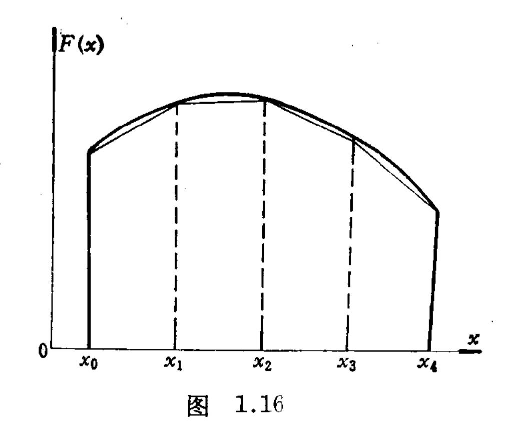


图 1.16

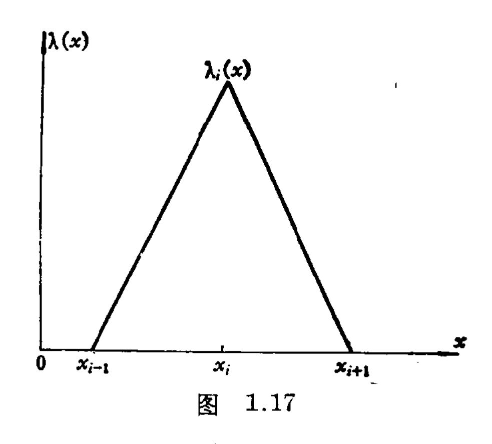


图 1.17

(x_{i+1}, f_{i+1})作抛物线(图1.18),

$$F(x)=\frac{(x-x_{i+1})(x-x_{i+\frac{1}{2}})}{(x_i-x_{i+1})(x_i-x_{i+\frac{1}{2}})}f_i+\frac{(x-x_i)(x-x_{i+1})}{(x_{i+\frac{1}{2}}-x_i)(x_{i+\frac{1}{2}}-x_{i+1})}f_{i+\frac{1}{2}}$$

$$+\frac{(x-x_i)(x-x_{i+\frac{1}{2}})}{(x_{i+1}-x_i)(x_{i+1}-x_{i+\frac{1}{2}})}f_{i+1} \tag{1.3.11}$$

$$x_i \leqslant x \leqslant x_{i+1},\ i=0,\ 1,\ \cdots,\ n-1$$

这是分段二次，总体为连续，但在 x_i 处导数有间断。这种插值也可以用基函数来表达，即

$$\left.\begin{aligned}\mu_i(x)&=2\lambda_i^2-\lambda_i\\ \mu_{i+\frac{1}{2}}(x)&=4\lambda_i\lambda_{i+1}\end{aligned}\right\} \tag{1.3.12}$$

如图1.19．这里 $\lambda_i=\lambda_i(x)$ 为分段线性插值的基函数(1.3.8)。注意 μ_i 在区间 $[x_{i-1}, x_{i+1}]$ 以外恒为 0，$\mu_{i+\frac{1}{2}}$ 在区间 $[x_i, x_{i+1}]$ 以外恒为 0。不难验证

$$\left.\begin{aligned}&\mu_i(x_j)=\delta_{ij},\quad \mu_i(x_{j+\frac{1}{2}})=0\\ &\mu_{i+\frac{1}{2}}(x_j)=0,\quad \mu_{i+\frac{1}{2}}(x_{j+\frac{1}{2}})=\delta_{ij}\end{aligned}\right\} \tag{1.3.13}$$

因此

$$F(x)=\sum_{i=0}^{n} f_i\mu_i(x)+\sum_{i=0}^{n-1} f_{i+\frac{1}{2}}\ \mu_{i+\frac{1}{2}}(x) \tag{1.3.14}$$

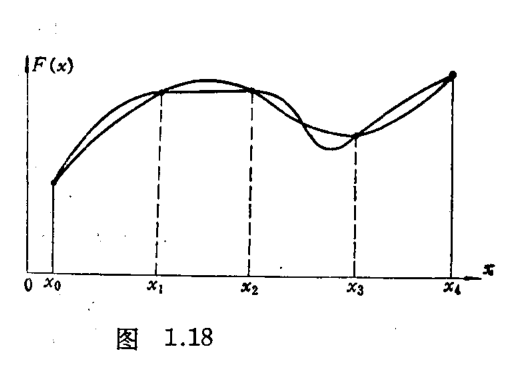


图 1.18

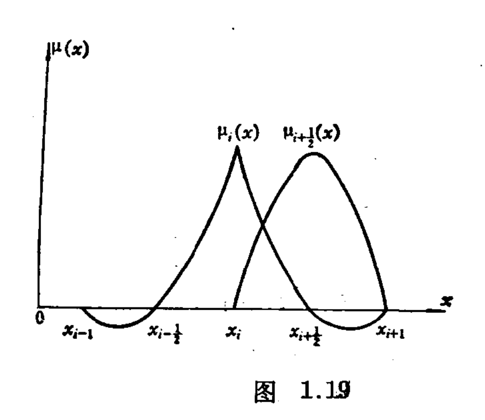


图 1.19

(四)分段三次埃尔米特插值

在每个子区间 $[x_i, x_{i+1}]$ 用其两端点的函数值及导数值 f_i, f'_i, f_{i+1}, f'_{i+1} 作三次埃尔米

特插值。因此是分段三次，总体地直至一阶导数连续，但二阶导数在 x_i 处有间断(图 1.20)。根据式(1.2.15)取插值基函数($i=0, 1, \cdots, n$)

$$
\alpha_i(x)=3\lambda_i^2-2\lambda_i^3,\quad a\leqslant x\leqslant b
$$

$$
\beta_i(x)=\begin{cases}-h_{i-\frac{1}{2}}\lambda_i, & a\leqslant x\leqslant x_i\\ h_{i+\frac{1}{2}}, & x_i\leqslant x\leqslant b\end{cases}\tag{1.3.15}
$$

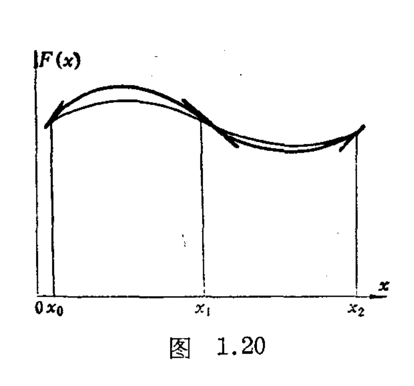


图 1.20

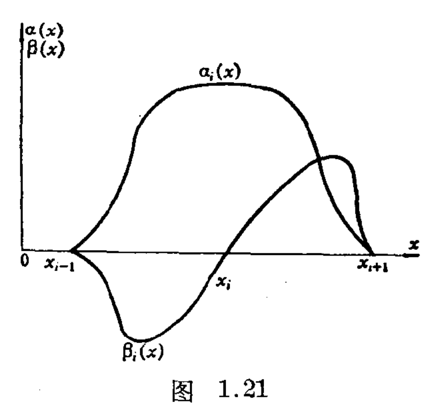


图 1.21

这里 $\lambda_i=\lambda_i(x)$ 就是分段线性插值的基函数(1.3.8)。注意 α_i, β_i 在区间 $[x_{i-1}, x_{i+1}]$ 以外恒为 0(图 1.21)。不难验证

$$
\alpha_i(x_j)=\delta_{ij},\quad \alpha_i'(x_j)=0,\quad \beta_i(x_j)=0,\quad \beta_i'(x_i)=\delta_{ij}\tag{1.3.16}
$$

因此所要求插值函数可以表为

$$
F(x)=f_i\alpha_i(x)+f_i'\beta_i(x)+f_{i+1}\alpha_{i+1}(x)+f_{i+1}'\beta_{i+1}(x)\tag{1.3.17}
$$

$$
x_i\leqslant x\leqslant x_{i+1},\quad i=0, 1, \cdots, n-1
$$

或者

$$
F(x)=\sum_{i=0}^{n}[f_i\alpha_i(x)+f_i'\beta_i(x)],\quad a\leqslant x\leqslant b\tag{1.3.18}
$$

关于稳定性和收敛性

上述几种插值法有一些共同之点。它们都是用一些低次插值多项式“装配”或拼凑起来的，在拼接点有一定连续性或光滑性。此外，在每个子区间上的插值函数只依赖于本区段上的一些特定节点值，而与其外的节点值无关。因此，都是“局部化”的。这也表现在基函数上，每个基函数只在一个局部范围内不为 0，出了这个范围就恒为 0。这样每个节点值只影响到直接衔接的一两个区段而不及以远。因此，节点的数据误差基本上不扩散、不放大，从而保证了当节点数 n 递增时插值过程的数值稳定性。

通常称一个函数为紧凑的，指它在某个有界区间以外恒为 0。紧凑函数只在有限的范围内活跃。以上所述的分段低次多项式插值的基函数都是紧凑的。由于插值基函数就是数据误差的影响函数，因此，这种紧凑性就保证了分段插值过程的稳定性。与此相反，在1.2.2节中的高次拉格朗日插值的基函数 $l_i(x)$ 则不是紧凑的，从而带来了插值的不稳定性。

关于收敛性，也可以统一地论证。上述四种情况，分别是分段 m 次插值，$m=0, 1, 2, 3$。$m=0$ 时，函数 $F(x)$ 本身有间断；$m>0$ 时，$F(x)$ 直至 $(m-1)$ 阶导数为连续，命

$$
M_{m+1}=\max_{a\leqslant x\leqslant b}|f^{(m+1)}(x)|\tag{1.3.19}
$$

$$h=\max(h_{\frac{1}{2}},\ h_{1+\frac{1}{2}},\ \cdots,\ h_{n-\frac{1}{2}}) \tag{1.3.20}$$

于是将余项估计(1.2.27)、(1.2.32)分别运用于每个子区间，并汇总，这样就得到

$$|f^{(p)}(x)-F^{(p)}(x)|\leqslant\frac{1}{(m+1-p)!}M_{m+1}h^{m+1-p} \tag{1.3.21}$$

$$a\leqslant x\leqslant b,\quad 0\leqslant p\leqslant m,\quad m=0,\ 1,\ 2,\ 3$$

因此，当 $h\to 0$ 时，相应的插值函数 $F(x)$ 连同其导数 $F^{(p)}_{(x)}$，一致收敛于原函数 $f(x)$ 及其导数 $f^{(p)}(x)$，$0\leqslant p\leqslant m$，$m=0,\ 1,\ 2,\ 3$。正如在1.2.3节中指出的，一般以整体的解析函数为基础的插值方法中，当节点加密时，插值过程可以不收敛，这就是说，插值与逐次逼近往往是两码事。但是在这里的分段插值法中，收敛性得到充分保证。当节点加密时的分段插值过程，同时也就是逼近过程，两者是统一的。

还应指出，在上述方法中，由于分段采用低次多项式，所以公式简单，并且避免了计算机上作高次乘幂时常遇到的上溢和下溢的困难，从而这也是有利的。

上面几种分段插值法的总体光滑度都不高，这对于某些应用是有缺陷的。在这种局部化的方法中，要增高光滑度就得采用更高阶的导数值，次数也相应提高。为了只用函数值本身，并在尽可能低的次数下达到较高的光滑度，可以采用所谓样条(spline)插值方法，详见下节。

1.3.2 三次样条插值

给定一组节点 $x_0<x_1<\cdots<x_{n-1}<x_n$ 及节点值 $f_0,\ f_1,\ \cdots,\ f_{n-1},\ f_n$，可以提出这样的问题：要求作一个插值函数 $F(x)$，它在每个子区间 $[x_0,\ x_1]$，$[x_1,\ x_2]$，$\cdots$，$[x_{n-1},\ x_n]$ 上分别都是三次多项式，而在整个区间 $[x_0,\ x_n]$ 上直至二阶导数为连续，并且满足插值条件

$$F(x_0)=f_0,\ F(x_1)=f_1,\ \cdots,\ F(x_n)=f_n \tag{1.3.22}$$

这里共有 n 个区段，每段三次多项式有四个系数，共有 $4n$ 个参数待定。为了保证 $F(x)$ 在整区间 $[x_0,\ x_n]$ 上直至二阶导数连续，只需要求 $F(x)$，$F'(x)$，$F''(x)$ 在内节点 $x_1,\ \cdots,\ x_{n-1}$ 连续，即

$$F(x_i-0)=F(x_i+0),\quad F'(x_i-0)=F'(x_i+0),$$

$$F''(x_i-0)=F''(x_i+0),\quad i=1,\ \cdots,\ n-1$$

这有 $3(n-1)$ 个条件，连同插值条件(1.3.22) $n+1$ 个，共可列出 $3(n-1)+n+1=4n-2$ 个条件。为了定出 $4n$ 个参数，还缺两个条件。因此，还需在左右两端各给一个"边界条件"，例如给定端点的一阶或二阶导数值，如

$$F'(x_0)=f'_0,\quad F'(x_n)=f'_n$$

或

$$F''(x_0)=f''_0,\quad F''(x_n)=f''_n$$

或其种种组合等。这样，在增补了两个边界条件后，问题可以唯一定解。这种插值叫做样条插值，导源于生产实践中用于放样的"样条"。当函数为分段 m 次多项式，在分段点直至 $m-1$ 阶导数为连续时，通称为 m 次样条函数，简称为样条。这里说的是三次样条，而在1.3.1节中所说的台阶状函数和折线状函数，则分别是零次和一次样条。

为了作三次样条插值，取 $F''(x_i)=M_i$，$i=0,\ 1,\ \cdots,\ n$ 作为待定参数比较方便。由于 $F(x)$ 为分段三次多项式，所以 $F''(x)$ 为分段线性，因此

$$F''(x)=\frac{x_{i+1}-x}{h_i}M_i+\frac{x-x_i}{h_i}M_{i+1} \tag{1.3.23}$$

$$h_i=x_{i+1}-x_i,\quad x_i\leqslant x\leqslant x_{i+1},\quad i=0,1,\cdots,n-1$$

在每个区间$[x_i, x_{i+1}]$上对此积分两次，利用两端条件$F(x_i)=f_i$, $F(x_{i+1})=f_{i+1}$即得

$$F'(x)=-\frac{(x_{i+1}-x)^2}{2h_i}M_i+\frac{(x-x_i)^2}{2h_i}M_{i+1}-\left(\frac{f_i}{h_i}-\frac{h_iM_i}{6}\right)+\left(\frac{f_{i+1}}{h_i}-\frac{h_iM_{i+1}}{6}\right) \tag{1.3.24}$$

$$F(x)=\frac{(x_{i+1}-x)^3}{6h_i}M_i+\frac{(x-x_i)^3}{6h_i}M_{i+1}+(x_{i+1}-x)\left(\frac{f_i}{h_i}-\frac{h_iM_i}{6}\right)+(x-x_i)\left(\frac{f_{i+1}}{h_i}-\frac{h_iM_{i+1}}{6}\right). \tag{1.3.25}$$

这时过点条件(1.3.22)，以及函数本身和二阶导数的连续均已保证了，还需要求在交接处x_i, $i=1,\cdots,n-1$，从左$[x_{i-1}, x_i]$及右$[x_i, x_{i+1}]$的一阶导数连续，

即
$$F'(x_i-0)=F'(x_i+0)$$

于是，可以排出$n-1$个方程

$$h_{i-1}M_{i-1}+2(h_{i-1}+h_i)M_i+h_iM_{i+1}=6\left[\frac{f_{i+1}-f_i}{h_i}-\frac{f_i-f_{i-1}}{h_{i-1}}\right] \tag{1.3.26}$$

$$i=1,2,\cdots,n-1$$

为了定出$n+1$个未知数$M_0, M_1, \cdots, M_n$还需附加两个条件。

附加条件的给法可以多种多样，通常是在两个端点x_0, x_n各给一个边界条件。例如，

(1) 给定导数值$F'(x_0)=f'_0$或$F'(x_n)=f'_n$.

即
$$f'_0=-\frac{h_0}{2}M_0-\left(\frac{f_0}{h_0}-\frac{h_0M_0}{6}\right)+\left(\frac{f_1}{h_0}-\frac{h_0M_1}{6}\right) \tag{1.3.27}$$

$$\left.\begin{aligned}&\text{左端 } x_0: && 2h_0M_0+h_0M_1=6\left[\frac{f_1-f_0}{h_0}-f'_0\right]\\&\text{右端 } x_n: && h_{n-1}M_{n-1}+2h_{n-1}M_n=6\left[f'_n-\frac{f_n-f_{n-1}}{h_{n-1}}\right]\end{aligned}\right\} \tag{1.3.28}$$

当曲线在端点的斜率很明确时，可以采用这种边界条件。例如当端点是一个局部极值点，这时应有水平斜率，则可给$f'=0$。

(2) 给定二阶导数值$F''(x_0)=f''_0$及$F''(x_n)=f''_n$:

$$\left.\begin{aligned}&\text{左端 } x_0: && M_0=f''_0\\&\text{右端 } x_n: && M_n=f''_n\end{aligned}\right\} \tag{1.3.29}$$

当曲线在端点的行为近似于反折点，即$f''=0$，则可给$M_0=0$及$M_n=0$。

(3) 一般的边界条件可以表成

$$\left.\begin{aligned}M_0&=\alpha_0M_1+\beta_0\\M_n&=\alpha_nM_{n-1}+\beta_n\end{aligned}\right\} \tag{1.3.30}$$

也就是适当选取常数α_0, β_0及α_n, β_n，使它们与曲线在端点的趋向相协调。一般当无其它动机时，可取$M_0=M_1$, $M_n=M_{n-1}$，这相当于端点的$f'''=0$。

也可以给出所谓周期性边界条件，即认为函数$F(x)$在$[x_0, x_n]$向两端周期性延拓，而保持直到二阶导数的连续性。为此，自然有$f_0=f_n$。并命$h_{-1}=h_{n-1}$, $M_n=M_0$，从而除方程组

(1.3.26)外，增加一个方程，表示 $F'(x_0-0)=F'(x_0+0)$，即

$$2(h_{n-1}+h_0)M_0+h_0M_1+h_{n-1}M_{n-1}=6\left[\frac{f_1-f_0}{h_0}-\frac{f_0-f_{n-1}}{h_{n-1}}\right] \tag{1.3.31}$$

这样，连同式(1.3.26)共有 n 个方程，以解 n 个未知数 $M_0, M_1, \cdots, M_{n-1}$。

周期性边界条件，可用于封闭曲线的插值。

对于一般的边界条件 1, 2, 3 待解的方程，可以表成

$$\begin{bmatrix} b_0 & c_0 & & & & 0 \\ a_1 & b_1 & c_1 & & & \\ & & \ddots & & & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & 0 & & & a_n & b_n \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix} \tag{1.3.32}$$

其中

$$a_i=h_{i-1},\ b_i=2(h_{i-1}+h_i),\ c_i=h_i,\quad d_i=6\left(\frac{f_{i+1}-f_i}{h_i}-\frac{f_i-f_{i-1}}{h_{i-1}}\right) \tag{1.3.33}$$

$$i=1,\ 2,\ \cdots,\ n-1$$

端点条件类别列表如下：

表 1.1

		1	2	3
左端	b_0	$2h_0$	1	1
	c_0	h_0	0	$-\alpha_0$
	d_0	$6\left(\frac{f_1-f_0}{h_0}-f'_0\right)$	f''_0	β_0
右端	a_0	h_{n-1}	0	$-\alpha_n$
	b_0	$2h_{n-1}$	1	1
	d_0	$6\left(f'_n-\frac{f_n-f_{n-1}}{h_{n-1}}\right)$	f''_n	β_n

这是对角元占优势的三对角线带状矩阵，可以用消元法，即追赶算法(详见第八章)求解如下：

命 $\qquad q_{-1}=0,\quad u_{-1}=0,\quad c_n=0$

对于 $\qquad k=0,\ 1,\ \cdots,\ n$

$$\left.\begin{aligned} p_k&=a_kq_{k-1}+b_k \\ q_k&=-c_k/p_k \\ u_k&=(d_k-a_ku_{k-1})/p_k \end{aligned}\right\} \tag{1.3.34}$$

这是正消过程。然后进行"回代"，

$$M_n=u_n$$

对于 $\qquad k=n-1,\ n-2,\ \cdots,\ 1,\ 0$

$$M_k=q_kM_{k+1}+u_k \tag{1.3.35}$$

对于周期性边界条件，由于 $M_0=M_n$，可以 $M_1, \cdots, M_n$ 为未知数而得方程组

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 & a_1 \\ a_2 & b_2 & c_2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & 0 \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ c_n & 0 & \cdots & 0 & a_n & b_n \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix} \tag{1.3.36}$$

这里 a_i, b_i, c_i, d_i $(i=1, 2, \cdots, n)$ 均由公式(1.3.33)给出，其中要用到

$$h_0=h_n, \quad f_0=f_n$$

这是循环型的三对角线带状矩阵。同样，可采用消去法。当正向消去 $n-1$ 步后得到等价于式(1.3.36)中前 $n-1$ 个方程的方程组

$$M_k=q_kM_{k+1}+s_kM_n+u_k, \quad k=1, 2, \cdots, n-1$$

比式(1.3.35)多出一项 s_kM_n。视 M_n 为参数，则上式又可表为

$$M_k=t_kM_n+v_k, \quad k=n-1, n-2, \cdots, 1 \tag{1.3.37}$$

系数 q_k, s_k, u_k 以及 t_k, v_k 分别满足正反递推关系：

命 $\qquad q_0=0, \ u_0=0, \ s_0=1$

对于 $\qquad k=1, 2, \cdots, n-1$

$$\left.\begin{aligned} p_k&=a_kq_{k-1}+b_k \\ q_k&=-c_k/p_k \\ u_k&=(d_k-a_ku_{k-1})/p_k \\ s_k&=-a_ks_{k-1}/p_k \end{aligned}\right\} \tag{1.3.38}$$

命 $\qquad t_n=1, \ v_n=0$

对于 $\qquad k=n-1, n-2, \cdots, 1$

$$t_k=q_kt_{k+1}+s_k$$

$$v_k=q_kv_{k+1}+u_k$$

参数 M_n 则可用方程组(1.3.36)中第 n 个方程

$$c_nM_1+a_nM_{n-1}+b_nM_n=d_n$$

来定，即

$$c_nM_1+a_n(t_{n-1}M_n+v_{n-1})+b_nM_n=d_n$$

因此

$$M_n=(d_n-c_nv_1-a_nv_{n-1})/(c_nt_1+a_nt_{n-1}+b_n) \tag{1.3.39}$$

然后用式(1.3.37)计算 M_k。这样，完整的计算公式依次是(1.3.38)，(1.3.39)，(1.3.37)。

1.3.3 参数表达的样条插值

当曲线的函数 $y=f(x)$ 有奇异性时，用自变量 x 的多项式形式的样条插值就不甚适应。但是，应该注意到，有时这种奇异性是由自变量的选取所引起的而不是曲线固有的。例如单位圆

$$x^2+y^2=1$$

本身并不具有任何奇异性，但将它表为

$$y=f(x)=\pm\sqrt{1-x^2}, \quad -1\leqslant x\leqslant 1$$

则首先它已经不是单值函数，而且在 $x=\pm 1$ 处有导数奇点 $y'=\pm\infty$，这就带来了插值处理的困难。一般当然是可以解决的，例如在插值函数中引进足以反映奇点行为的成分，但方法上复杂化了。但是如果用极坐标 r, θ 来表达，则成为

$$r=r(\theta)\equiv 1$$

或者用参数表达

$$\begin{cases} x=\cos\theta \\ y=\sin\theta \end{cases}$$

就可以避免上述困难。为了面向一般的问题，我们讨论参数表达的平面曲线：

$$x=x(\alpha), \quad y=y(\alpha) \tag{1.3.40}$$

特别有利的是以弧长作为参数 α，因为这是曲线自身的内在坐标。

设待插曲线的样点为

$$p_i=(x_i, y_i), \quad i=0, 1, \cdots, n$$

由于弧长本身还是不知道的，因此可以采用"积累弦长"：

命 $$s_0=0$$

$$s_i=s_{i-1}+\sqrt{(x_{i-1}-x_i)^2+(y_{i-1}-y_i)^2}$$

$$i=1, 2, \cdots, n \tag{1.3.41}$$

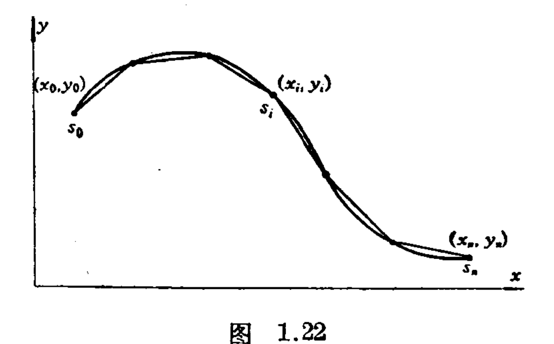


图 1.22

设想以积累弦长 s 为自变量，于是两个函数 $x(s)$ 及 $y(s)$ 在节点 $s=s_0, s_1, \cdots, s_n$ 的样点值为(图 1.22)

$$\left.\begin{aligned} x(s_i)=x_i \\ y(s_i)=y_i \end{aligned}\right\} i=0, 1, \cdots, n \tag{1.3.42}$$

在区间 $s_0\leqslant s\leqslant s_n$ 上分别作样条插值函数 $X(s)$, $Y(s)$，即得过点插值曲线

$$P(s)=(X(s), Y(s))$$

对于封闭曲线，则命 $(x_0, y_0)=(x_n, y_n)$ 而采用周期性边界条件(图 1.23)。例如对单位圆 $x^2+y^2=1$ 取等分节点(图 1.24)的计算结果如下：

节 点 个 数 n	计算所得半径的最大误差
4	$\leqslant 0.01$
8	$\leqslant 0.00112$
12	$\leqslant 0.000165$

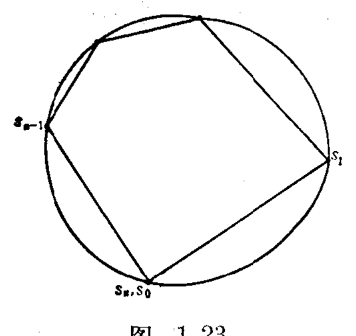
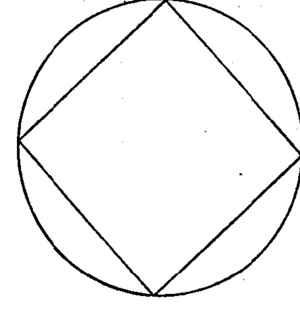


图 1.23


图 1.24

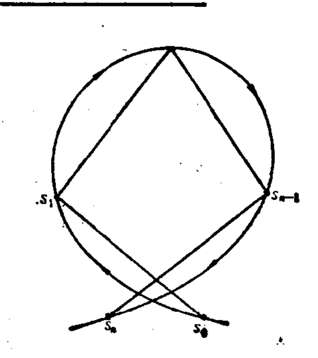


图 1.25

这个方法的优点是,它采用了某种内在的坐标,因而不依赖于曲线的形状走向。它不仅对于封闭曲线，而且对于更一般的如自相交曲线(图 1.25)也适用，因此有很大的通用性。通常工程绘图用的曲线板所含的几何信息,用约 30 个节点值就足以表达。这个方法显然很容易推广到空间曲线

$$x=x(s),\ y=y(s),\ z=z(s)$$

的插值。

1.3.4 样条插值的物理背景

样条插值直接导源于生产实践,它有明确的物理背景。

以一些集中力作用于张紧的弦线,则相邻着力点之间形成直线,仅坡度有间断。因此勒过一些样点的张紧弦线,就自动形成了折线,即一次样条插值。木工用墨斗墨绳,成衣工人用粉袋弹线以进行放样,就是这个原理。从数学上看,在小变形时,弹性弦的平衡方程是 $y''=q$。$y=y(x)$ 为弹性变形；$q=q(x)$ 为载荷分布。当载荷为作用于节点 $x_0, x_1, \cdots, x_n$ 的集中力时，

$$q(x)=\sum_{i=0}^{n} q_i\delta(x-x_i),\quad \delta(x-x_i)$$

为集中于点 $x=x_i$ 的“脉冲函数”,即 δ 函数(见第三章§3.2)。于是

$$y''(x)=\sum_{i=0}^{n} q_i\delta(x-x_i) \tag{1.3.43}$$

因此,在相邻着力点之间,$x_i\leqslant x\leqslant x_{i+1}$, $y''(x)\equiv 0$, 即 y 为一次多项式。在着力点 $x=x_i$ 上 y'' 为脉冲状间断，即 y' 为台阶状间断而 y 本身连续。因此整体地 $y(x)$ 为一次样条，即折线函数。

类似地，以一些集中力作用弹性薄条(可以看成弹性梁，则该条自动弯曲形成光滑的曲线，仅在着力点有三阶导数的间断，而坡度、曲率都是连续的。这是因为小变形时的弹性梁的平衡方程为 $y^{(4)}=q$,而当 q 为作用于节点 $x_0, x_1, \cdots, x_n$ 的集中力时,则成为

$$y^{(4)}(x)=\sum_{i=0}^{n} q_i\delta(x-x_i) \tag{1.3.44}$$

与式(1.3.43)相仿,只是左端二阶导数代成四阶导数。这时,在相邻着力点之间 $x_i\leqslant x\leqslant x_{i+1}$, $y^{(4)}(x)\equiv 0$, 即 y 为三次多项式。而在着力点 $x=x_i$ 上，$y^{(4)}$ 为脉冲状间断，即 $y^{(3)}$ 为台阶状间断，$y^{(2)}$, $y^{(1)}$, y 就都连续,因此 $y(x)$ 为三次样条函数。在实践中就有这种情况,如用木条或薄钢条或其它弹性材料做“样条”,并用压铁压住,以强使它通过一些样点而自动形成光顺的插值曲线。机械工人就是运用这一科学原理来进行放样的。数学上的样条插值方法正是模拟这一原理而发展起来的。

当弹性体达到弹性平衡时,应变能必定达到极小,反之亦然,这就是所谓最小势能原理。在小变形时，弹性弦和梁的应变能分别表为 $\int_a^b (y')^2dx$, $\int_a^b (y'')^2dx$。既然一次及三次样条是与弦及梁连系着的,则它们也必然具有相应的极值性质。

可以证明,在区间 $x_0\leqslant x\leqslant x_n$ 上,在一切满足

$$F(x_i)=f_i,\quad i=0,\ 1,\ \cdots,\ n \tag{1.3.45}$$

的连续函数中，使得积分 $\int_{x_0}^{x_n} (F'(x))^2dx$ 达到极小的函数就是以式(1.3.45)为条件的一次样条(折线)插值函数。类似地,在一切满足式(1.3.45)并且具有连续二阶导数的函数中,使

得积分 $\int_{x_0}^{x_n}(F''(x))^2dx$ 达到极小的函数就是在式(1.3.45)连同边界条件

$$F''(x_0)=F''(x_n)=0$$

下的三次样条插值函数。在这个意义下，三次样条插值可以说是在各种可能的插值中使得均方曲率(曲率$=y''(1+y'^2)^{-3/2}\approx y''$，当 $y'\approx 0$ 时)为最小，即在一定意义下最为"光顺"。

样条插值也有良好的收敛性。关于一次样条问题已由式(1.3.21)说明。关于三次样条，当原函数 $f(x)$ 足够光滑，例如，具有连续的四阶导数时，则可以证明，当最长间距 $h\to 0$ 时，$F(x)$ 以及 $F'(x)$ 均以 $O(h^2)$ 的速度一致收敛于 $f(x)$ 及 $f'(x)$，而 $f''(x)$ 以 $O(h)$ 的速度一致收敛于 $f''(x)$。由于这种带导数的一致收敛性以及"最光顺"性，因此三次样条插值也可以作为数值微分的工具(见 §1.4)。

1.3.5 方法比较

样条插值对比于其它的分段插值的主要优点是，保证了直到二阶导数的连续性，因此光滑度较高。直观地看，一条曲线如有间断或坡度的间断则是很显眼的，对于曲率的间断则要仔细打量后才能察觉。至于三阶导数的间断，则肉眼就很难辨认了。样条插值比其他的片段插值提高了光滑度，因为它达到了二阶导数连续，仅在节点处有三阶导数的间断。这是样条插值法的主要优点。

前面介绍的那些较低阶的片段插值都是局部化的，即每个节点只影响到附近少数几个间距，从而带来了计算上的方便，可以步进地进行，即从一端开始按显式一步一步地插过去。同时也带来了内在的高度稳定性。

样条插值则不是局部化的，每个节点影响到全局。因此计算不方便。它是隐式的，即需要联解一个代数方程，在样点数量很大时很不利。与此同时，稳定性也就较差于那些局部化的方法。但是，样条节点的影响是随着远离该点而衰退的，因此它的稳定性对比于高次多项式插值要好得多，但比低阶片段插值为差。但是由于存在着误差的远距离的扩散，使得样条插值也会有"多余"的波动，特别在间距不均匀以及其它一些特殊场合更为显著。图1.26表示一个不利的情况，样点从0至7指数状单调上升，从点7起取常值。样条插值的结果在点7至8之间出现不合理的隆起。

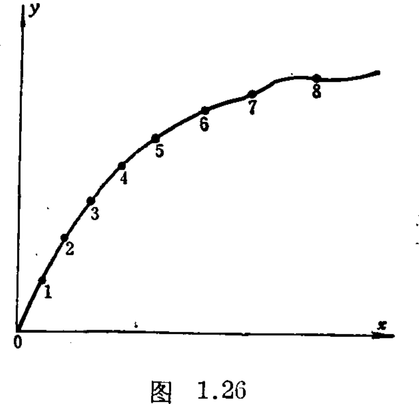


图 1.26

§1.4 数值微分

设给定了某个变量 $f(t)$ 的一批样点值 $f_0=f(t_0)$，$f_1=f(t_1)$，…，譬如说某移动目标的观测值，要求据此推算一阶、二阶导数 $f'(t)$，$f''(t)$ 在各个时刻的值，即是速度、加速度，这就是数值微分的问题。

1.4.1 数值微分公式

在微积分里，微商是极限的概念

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}=\lim_{h\to 0}\frac{f(x)-f(x-h)}{h}=\lim_{h\to 0}\frac{f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right)}{h} \tag{1.4.1}$$

显然，取其达到极限以前的形式就得到微商的差商近似

$$f'(x)\approx\frac{f(x+h)-f(x)}{h}\approx\frac{f(x)-f(x-h)}{h}\approx\frac{f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right)}{h} \tag{1.4.2}$$

从几何上看就相当于用弧段的内接弦的斜率代替切线的斜率(图 1.27)。

用节点值来表示时：

$$\frac{f_{i+1}-f_i}{x_{i+1}-x_i}=\frac{f_{i+1}-f_i}{h}\approx\begin{cases}f'(x_i) & (1.4.3)\\ f'(x_{i+1}) & (1.4.4)\\ f'(x_{i+\frac{1}{2}}) & (1.4.5)\end{cases}$$

$$x_{i+1/2}=\frac{1}{2}(x_i+x_{i+1})$$

图 1.27

从几何上看，弧段的内接弦的斜率与切线斜率的平行程度在中点优于两端点(图 1.27)。也可以分别在 x_i, x_{i+1}, $x_{i+1/2}$ 作幂次展开而得截断误差即差商与微商之差

$$E=\begin{cases}\frac{h}{2}f''(\xi)\approx O(h) & (1.4.3)'\\ -\frac{h}{2}f''(\xi)\approx O(h) & (1.4.4)'\\ \frac{h^2}{24}f'''(\xi)\approx O(h^2) & (1.4.5)'\end{cases}$$

因此，用两点差分作为其中点处的导数值，即公式

$$\frac{f_{i+1}-f_i}{h}=\frac{f_{i+1}-f_i}{x_{i+1}-x_i}\approx f'(x_{i+1/2}),\quad E=O(h^2) \tag{1.4.6}$$

精度最好，它提高了一阶。这是数值微分的基本公式，据此可以推出绝大多数实用的数值微分公式。

对于等距节点，如果要求不是在半点上而是在整点上的导数值，则可以按基本公式(1.4.6)求出 $x_{i-1/2}$, $x_{i+1/2}$ 处导数值后再线性插到 x_i 得

$$f'(x_i)\approx\frac{1}{2}(f'_{i+1/2}+f'_{i-1/2})\approx\frac{1}{2}\left(\frac{f_{i+1}-f_i}{h}+\frac{f_i-f_{i-1}}{h}\right)=\frac{f_{i+1}-f_{i-1}}{2h}\quad E=O(h^2) \tag{1.4.7}$$

事实上也可以直接从基本公式(1.4.6)得到，只是把间距扩大一倍。

对于二阶导数则有

$$f''(x_i)\approx\frac{f'_{i+1/2}-f'_{i-1/2}}{h}\approx\frac{\frac{f_{i+1}-f_i}{h}-\frac{f_i-f_{i-1}}{h}}{h}=\frac{1}{h^2}(f_{i+1}-2f_i+f_{i-1}),\quad E=O(h^2) \tag{1.4.8}$$

类似地，可得

$$f'''(x_{i+1/2})\approx\frac{1}{h^3}(f_{i+2}-3f_{i+1}+3f_i-f_{i-1}),\quad E=O(h^2) \tag{1.4.9}$$

$$f''''(x_i)\approx\frac{1}{h^4}(f_{i+2}-4f_{i+1}+6f_i-4f_{i-1}+f_{i-2}),\quad E=O(h^2) \tag{1.4.10}$$

对于不等距节点，一种形式是，原则上先作相应节点的插值多项式，对此作解析的微分，即可得到所需点的各阶导数值。设有一列节点

$$x_i<x_{i+1}<\cdots<x_{i+m}$$

$$f_j=f(x_j),\ j=i,\ i+1,\ \cdots,\ i+m$$

作 m 次插值多项式

$$F_i(x)=\sum_{j=i}^{i+m}f_jl_{i,j}(x),\quad l_{i,j}(x)=\prod_{\substack{k=i\\k\neq j}}^{i+m}\frac{x-x_k}{x_j-x_k} \tag{1.4.11}$$

于是在点 $x=\xi$ 处的 n 阶($n\leqslant m$)数值微商即差商公式就是

$$f^{(n)}(\xi)\approx F_i^{(n)}(\xi)=\sum_{j=i}^{i+m}f_jl_{i,j}^{(n)}(\xi) \tag{1.4.12}$$

特别当 $m=n$ 时

$$l_{i,j}^{(m)}(x)\equiv m!\prod_{\substack{k=i\\k\neq j}}^{i+m}\frac{1}{x_j-x_k}=\alpha_{i,j} \tag{1.4.13}$$

为常数，因得 m 阶数值微商公式

$$f^{(m)}(\xi)\approx\sum_{j=i}^{i+m}\alpha_{i,j}f_j \tag{1.4.14}$$

这时，不论 ξ 为何，结果是相同的，但以 ξ 取在节点 $x_i,\ \cdots,\ x_{i+m}$ 的中点时比较精确，于是

$$f^{(m)}(x_{i+\frac{m}{2}})\approx\sum_{j=i}^{i+m}\alpha_{i,j}f_j \tag{1.4.15}$$

当 $m=2q=$ 偶数时，$x_{i+\frac{m}{2}}=x_{i+q}$，当 $m=2q+1=$ 奇数时，$x_{i+\frac{m}{2}}=x_{i+q+\frac{1}{2}}$ 了解为"半点"$\frac{1}{2}(x_{i+q}+x_{i+q+1})$。

如果命

$$\omega_i(x)=\prod_{k=i}^{i+m}(x-x_k) \tag{1.4.16}$$

则有(见 1.2.2 节)

$$\omega_i'(x_j)=\prod_{\substack{k=i\\k\neq j}}^{i+m}(x_j-x_k) \tag{1.4.17}$$

于是 $\alpha_{i,j}$ 可以表为

$$\alpha_{i,j}=\frac{m!}{\omega_i'(x_j)} \tag{1.4.18}$$

不难验证，在等间距 $x_{j+1}-x_j\equiv h$ 时

$$\alpha_{i,i+k}=\frac{(-1)^{m-k}}{h^m}C_k^m,\quad C_k^m=\frac{m!}{k!(m-k)!} \tag{1.4.19}$$

因此

$$f^{(m)}{}_{i+\frac{m}{2}}\approx\frac{1}{h^m}\sum_{k=0}^{m}(-1)^{m-k}C_k^mf_{i+k} \tag{1.4.20}$$

当 $m=1,\ 2,\ 3,\ 4$ 时前面已经导出。

1.4.2 数据误差对于微分的影响

从数值微分的截差公式看来，间距 h 愈小，精度愈高。但是在实际计算时问题远不是那样简单。例如取五位指数函数 $f(x)=e^x$ 表(表 1.2)[5]，用中心差商 $\frac{f_{i+1}-f_{i-1}}{2h}$ 来计算 $f'(1)$，其真值为 $e=2.7183$。分别取 $h=0.2$, 0.1 及 0.01 的结果及误差列在表 1.3。可以看到，当 h 从 0.2 缩小到 0.1 时确实得到改进，但 h 进一步缩至 0.01 时，则结果反而恶化。按 $\frac{1}{h^2}(f_{i+1}-2f_i+f_{i-1})$ 计算 $f''(1)$ (真值也是 $e=2.7183$) 时，用同样三种步长的结果列在表 1.4，其情况就更为突出。当 h 从 0.1 缩至 0.01 时，不仅没有提高精度，反而把有效数字丢光，连最高位数字也不对(参考[2]，第五章)。

表 1.2

x	e^x	……
0.00	1.0000	⋮
⋮	⋮	
0.90	2.4596	
⋮	⋮	
0.99	2.6912	
1.00	2.7183	
1.01	2.7456	
⋮	⋮	
1.10	3.0042	
2.00	7.3891	

表 1.3

h	$f'(e)=\frac{f(1+h)-f(1-h)}{2h}$	误　差
1.0	3.1946	−0.4736
0.1	2.723	−0.0047
0.01	2.71	−0.0083

表 1.4

h	$f''(1)=\frac{1}{h^2}(f(1+h)-2f(1)+f(1-h))$	误　差
1.0	2.9525	0.2342
0.1	2.72	0.0017
0.01	2	−0.7183

问题的根源在于，截断误差只是计算误差的一个部分，另一个部分是由原始数据的误差带来的。原始数据含有舍入误差或其它来源的误差是不可避免的，而差商的运算恰恰对此特别敏感，它随 h 的缩小而增大，即具有不稳定性。例如，各个样点数据的最大误差为 ε，而且由于随机性在相邻点可以反号，从而对于差商 $\frac{f_{i+1}-f_{i-1}}{h}$ 带来的误差在不利的情况下就达到 $\frac{\varepsilon+\varepsilon}{h}=\frac{2\varepsilon}{h}$。因此，误差随 h 的缩小而被放大。对于二阶差商 $\frac{f_{i+1}-2f_i+f_{i-1}}{h^2}$ 则带来误差 $\frac{(1+2+1)\varepsilon}{h^2}=\frac{4\varepsilon}{h^2}$，正如已经看到的，误差放大更甚。为了对比，可看看下述情况：在数值积分时，例如 $\int_{x_0} f dx \approx h[f_0+f_1+\cdots]$，其个别节点误差 ε 的影响为 $h\cdot\varepsilon$，这说明其影响被缩小了，因此是稳定的。因为微分计算误差是截断误差和数据误差两项的迭加，当步长缩减时前者减小，后者增大，因此，要缩小误差的影响，并不是一味缩小步长所能奏效的，这里有一个最优步长选取的问题。

我们将对数值微商中数据误差(主要考虑舍入误差)和截断误差与步长之间的关系作初步的分析，并试图估计最优步长。命真值 $f(x_i)=f_i$, $f'(x_i)=f'_i$, $f''(x_i)=f''_i$；并命 f_i^* 为 f_i 带有舍入的实际表达值。对于舍入，最好用相对误差来表示，因此将 f_i^* 表为

$$f_i^*=f_i(1+\delta_i)=f_i+f_i\delta_i \tag{1.4.21}$$

设相对误差的界限为 δ, $|\delta_i|\leqslant\delta$。当取 s 位有效数字(即字长取 s 位)时 $\delta\approx10^{-s}$。命 M_p 表示 $|f^{(p)}|$ 在有关区段上的上界, $|f_i|\leqslant M_0$, $|f^{(p)}(x)|\leqslant M_p$。

数值微商实际上是对 f_i^* 进行的。以一阶中心差商为例,总的误差是

$$\frac{1}{2h}(f_{i+1}^*-f_{i-1}^*)-f_i'=\left[\frac{1}{2h}(f_{i+1}-f_{i-1})-f_i'\right]+\left[\frac{1}{2h}(f_{i+1}\delta_{i+1}-f_{i-1}\delta_{i-1})\right] \quad (1.4.22)$$

右端两项分别就是截断和舍入误差。对于截差,利用幂次展开得到

$$\left|\frac{1}{2h}(f_{i+1}-f_{i-1})-f_i'\right|=\left|\frac{h^2}{6}f'''(\xi)\right|\leqslant\frac{M_3h^2}{6}=E_1(h)$$

至于舍入误差则有

$$\left|\frac{1}{2h}(f_{i+1}\delta_{i+1}-f_{i-1}\delta_{i-1})\right|\leqslant\frac{1}{2h}(|f_{i+1}\delta_{i+1}|-|f_{i-1}\delta_{i-1}|)\leqslant\frac{1+1}{2h}M_0\delta=\frac{M_0\delta}{h}=E_2(h)$$

于是

$$\left|\frac{1}{2h}(f_{i+1}^*-f_{i-1}^*)-f_i'\right|\leqslant\frac{M_3h^2}{6}+\frac{M_0\delta}{h}=E_1(h)+E_2(h)=E(h) \quad (1.4.23)$$

右端 $E(h)$ 虽只是误差的上界,但由于舍入误差在数值上和符号上的随机性,这个界限是可以达到的。因此,函数 $E(h)$ 基本上能反映误差的规律性。当 $h\to0$ 时,$E(h)\to\infty$,其舍入误差为主导;当 $h\to\infty$ 时 $E(h)\to\infty$,其截断误差为主导;在 $0\sim\infty$ 中间,$E(h)$ 有一个极小点 $h=h^*$。为此只需解方程 $E'(h)=\frac{M_3h}{3}-\frac{M_0\delta}{h^2}=0$,从而得到

$$h^*=\left(\frac{3M_0}{M_3}\right)^{\frac{1}{3}}\delta^{\frac{1}{3}}\approx O(\delta^{\frac{1}{3}}) \quad (1.4.24)$$

当步长 h 取为 h^* 时,数值微商的误差为最小,如取 $h<h^*$,则由于舍入的影响而误差反而增大,因此 h^* 可以作为步长选取的下界。

对于二阶差商公式,类似可以得到估计式

$$\left|\frac{1}{h^2}(f_{i-1}^*-2f_i^*+f_{i+1}^*)-f_i''\right|\leqslant\frac{M_4h^2}{12}+\frac{4M_0\delta}{h^2}=E_1(h)+E_2(h)=E(h) \quad (1.4.25)$$

$$h^*=\left(\frac{48M_0}{M_4}\right)^{\frac{1}{4}}\delta^{\frac{1}{4}}\approx O(\delta^{\frac{1}{4}}) \quad (1.4.26)$$

再回到前面举的实例。对表 1.2 取五位数字,末位上有舍入,故 $\delta\approx10^{-5}$; $f=f'=f''=\cdots=e^x$,故可取 $M_0/M_3=M_0/M_4=1$。根据式(1.4.24)和(1.4.26)可以估出

一阶差商: $h^*\approx3^{\frac{1}{3}}10^{-\frac{5}{3}}\approx0.07$

二阶差商: $h^*\approx(48)^{\frac{1}{4}}10^{-\frac{5}{4}}\approx0.15$

与表 1.3、表 1.4 的实况相近。

另一个较为直观的方法,是认为截断误差 E_1 与舍入误差 E_2 达到相同量级

$$E_1(h)\approx E_2(h)$$

时给出最优的即临界步长 h^*。据此,对于二阶中心差商而言,得结果与式(1.4.24)同;对一阶中心差商而言,则得

$$h^*=\left(\frac{6M_0}{M_3}\right)^{\frac{1}{3}}\delta^{\frac{1}{3}}\approx1.26\left(\frac{3M_0}{M_3}\right)^{\frac{1}{3}}\delta^{\frac{1}{3}}\approx O(\delta^{\frac{1}{3}})$$

其结果与(1.4.26)基本一致。

微分和积分运算在稳定性上的原则差别，从下面的解析例子(参考[1]的第二章)可以更清楚地看到。

设$f(x)$在图1.28中是实线，迭加了初始误差后，在图中如虚线所示。其积分$g(x)=\int_0^x f(x)dx$及微分$f'(x)$分别如图1.29及图1.30所示。其中，实线表示本来的函数，虚线表示迭加了误差的结果。

可以见到，初始的误差在积分过程中基本上被"吸收"，而在微分过程中被恶性放大。

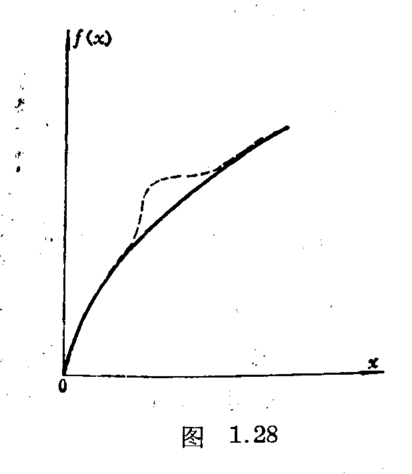


图 1.28

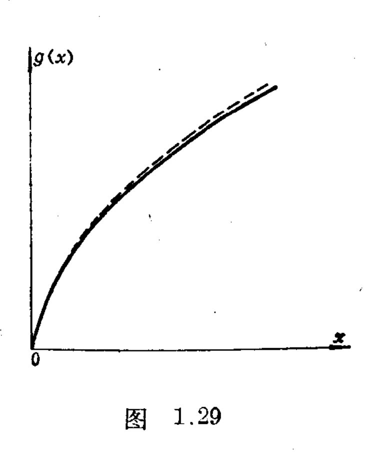


图 1.29

在上面所举的微分计算实例中，原始数据是数学函数表，只有舍入误差，而且被控制在第五位。这还是比较有利的情况。在一般的情况下，特别当初始数据是从实验得来时，数据误差达到更高的基准，数值微分也就更困难。由于数值微分的不可靠性，因此，一个处理原则是尽可能地避免它。例如，杆件的弹性纵振动可以表为下列二阶导数波动方程：

$$w_{tt}=w_{xx}$$

式中w为弹性位移；w_x为应力。数值解出波动方程，得出在离散节点上的w。为了求应力w_x，则需进行数值微分。也可以改变问题提法，例如引进函数$u=w_x$，$v=w_t$，而得含一阶导数的波动方程组

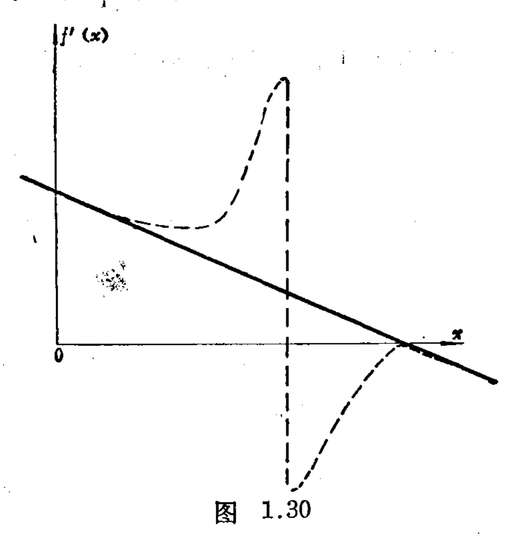


图 1.30

$$\begin{cases} u_t=v_x \\ v_t=-u_x \end{cases}$$

直接解出在节点上的v，这就是所要求的应力，而无需作数值微分。这在相当的计算代价下可以得到更为可靠的结果。

当离散数据来自实验或观测时，不可避免含有随机性误差，并服从一定的概率分布。这时数值微分的处理原则是，应该考虑到数据的整体，即必须经过平滑化以滤去随机性误差，然后在这个基础上进行微分。一个处理方法是，先用最小二乘法拟合一个较低次的逼近多项式(见§2.6)，然后对此微分。

此外，样条插值也可以作为数值微分的"工具"。这是因为，样条插值除了函数的收敛性(当步长缩小时)外，还保证导数的收敛性，并且这种插值在每一点的值还用到了数据的整体

的，含有平滑化的因素。为了求微分，可以先作样条插值(§2.2)，然后按公式(1.3.24)给出导数值。对于二阶导数，则可以先通过样条插值取得一阶导数，然后对此一阶导数作样条插值，再取其一阶导数作为原函数的二阶导数。

§1.5 一般样条和基样条

上面看到，分段多项式由于它的灵活性和稳定性，用处很大。当分段多项式每段次数 $\leqslant p$，在分段点即节点上直至 $p-1$ 阶导数连续时叫做 p 次样条，这时 p 阶导数可能有间断。零次样条是指分段为常数的函数，函数本身可能有间断。这就是1.3.1节中的分段零次插值，即台阶状函数。在1.3.1节中的分段线性插值即折线状函数便是一次样条。对这两种情况给出一些特定的样条作为基函数 $\varphi_i(x)$，使得同次的样条都可以表为它们的线性组合 $\sum a_i\varphi_i(x)$。在 §1.3.1 中所述的分段二次多项式，在节点上函数连续，一阶导数有间断，按照这里的意义不是严格的二次样条。在§1.3.2中给出了三次样条的插值法，但没有给出基函数的表达式。本节中将讨论一般的 p 次样条，并给出几种基函数的表达方式，使得一般样条可以用基样条为“元件”“装配”起来。这对于插值和拟合都是有利的。

1.5.1 一些简单的样条函数

样条函数的特点是光滑和灵活。p 次样条直至 $p-1$ 阶导数连续，这保证了它的光滑性；而 p 阶导数的间断则使样条有了转折自如的灵活性。一般的 p 次多项式也是一种 p 次样条。但其 p 阶导数连续(是常数)，因此它还只是名义上的样条。把 p 次幂 x^p 切去 $x<0$ 的半枝得到所谓“截断幂”

$$x_+^p=\begin{cases}0, & x<0\\ x^p, & x>0\end{cases} \tag{1.5.1}$$

这种手术在原点 $x=0$ 造成 p 阶导数间断，但仍保持其下各阶导数连续，这就得到最简单的“真正”的 p 次样条。

零次截断幂就是单位台阶函数

$$x_+^0=\begin{cases}0, & x<0\\ 1, & x>0\end{cases}$$

在 $x=0$ 处间断，跃值● 为 1。一次截断幂

$$x_+^1=\begin{cases}0, & x<0\\ x, & x>0\end{cases},\quad (x_+^1)'=\begin{Bmatrix}0, & x<0\\ 1, & x>0\end{Bmatrix}=x_+^0$$

函数连续，导数在 $x=0$ 处间断，跃值为 1。一般地，

$$(x_+^p)'=px_+^{p-1},\ (x_+^p)''=p(p-1)x_+^{p-2},\ \cdots,\ (x_+^p)^{(p-1)}=p!x_+^1,\ (x_+^p)^{(p)}=p!x_+^0$$

故 x_+^p 直至 $p-1$ 阶导数连续，p 阶导数在 $x=0$ 处间断，跃值为 $p!$。$p=0, 1, 2, 3$ 的曲线见图1.31的上列。

类似地，也可以把 p 次幂 x^p 的在 $x<0$ 的半枝翻个身，得到所谓“反转幂”

$$x_*^p=\begin{Bmatrix}-x^p, & x<0\\ x^p, & x>0\end{Bmatrix}=x^p\operatorname{sign}x=\begin{cases}|x|^p\operatorname{sign}x, & p=\text{偶数}\\ |x|^p, & p=\text{奇数}\end{cases} \tag{1.5.2}$$

● 当函数 $f(x)$ 在 $x=a$ 处间断时，该点的跃值是指从左和从右极限值的差数 $f(a+0)-f(a-0)$，也记作 $[f]_{x=a}$。当函数在该点连续时，跃值为 0。

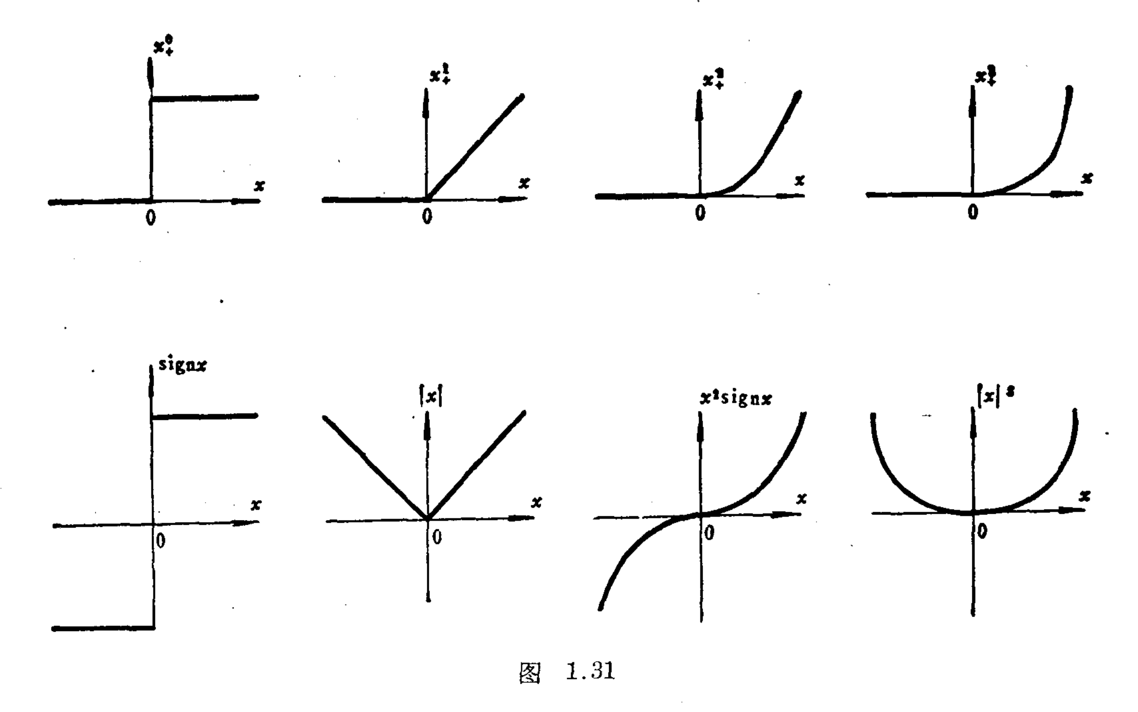


图 1.31

这种手术也在原点 $x=0$ 造成 p 阶导数间断而保持其下各阶导数的连续性，这也是真正的 p 次样条。零次反转幂就是符号函数

$$x_*^0=\operatorname{sign}x=\begin{cases}-1, & x<0\\ +1, & x>0\end{cases}$$

函数在 $x=0$ 处间断，跃值为 2。一次反转幂就是绝对值

$$x_*^1=|x|=\begin{cases}-x, & x<0\\ +x, & x>0\end{cases},\quad (x_*^1)'=\left\{\begin{matrix}-1, & x<0\\ 1, & x>0\end{matrix}\right\}=x_*^0$$

函数连续，导数在 $x=0$ 处间断，跃值为 2。一般地

$$(x_*^p)'=px_*^{p-1},\ (x_*^p)''=p(p-1)x_*^{p-2},\ \cdots,\ (x_*^p)^{(p-1)}=p!x_*^1,\ (x_*^p)^{(p)}=p!x_*^0$$

x_*^p 直至 $p-1$ 阶导数连续，p 阶导数在 $x=0$ 处间断，跃值为 $2p!$，$p=0, 1, 2, 3$ 的曲线见图 1.31 的下列。

不难验证

$$x_+^p=\frac{1}{2}(x_*^p+x^p) \tag{1.5.3}$$

$$x_*^p=x_+^p+(-1)^p(-x)_+^p \tag{1.5.4}$$

样条函数经过一些初等变换后仍保持为样条。例如，设 $f(x)$ 为 p 次样条，节点为 x_0, x_1, $\cdots$，将 $f(x)$ 平移一个距离 a 得函数 $f(x-a)$，则后者显然还是 p 次样条，其节点移至 x_0+a, x_1+a, $\cdots$。如对 $f(x)$ 作尺度变换而变为 $f(ax)$，则它也是 p 次样条，但节点变为 $\frac{x_0}{a}$, $\frac{x_1}{a}$, $\cdots$。设 $g(x)$ 是另一个 p 次样条，α, β 为常数，则线性组合 $\alpha f(x)+\beta g(x)$ 也是 p 次样条，其节点为 $f(x)$ 的节点加上 $g(x)$ 的节点。p 次样条 $f(x)$ 的导数 $f'(x)$ 是 $p-1$ 次样条，节点相同。利用这些性质可以从一些最简单的样条如截断幂或反转幂出发构成一系列新的样条。特别是从截断幂或反转幂通过“差分”构造出具有紧凑特点的所谓山丘形样条。

定义一阶差分

$$\Delta^1 f(x)=f\left(x+\frac{1}{2}\right)-f\left(x-\frac{1}{2}\right) \tag{1.5.5}$$

它是把函数 $f(x)$ 左、右移 $\frac{1}{2}$ 再相减而成的。显然，当 $f(x)$ 为 p 次样条时，$\Delta^1 f(x)$ 也是 p 次样条，不过节点有所不同。

取 0 次截断幂 x_+^0 的一阶差分，

$$\Delta^1 x_+^0=\left(x+\frac{1}{2}\right)_+^0-\left(x-\frac{1}{2}\right)_+^0=\begin{cases}0-0=0, & x<-\frac{1}{2}\\ 1-0=1, & -\frac{1}{2}<x<\frac{1}{2}\\ 1-1=0, & \frac{1}{2}<x\end{cases}$$

这是矩形的台阶状函数(1.3.1 节)，是 0 次样条，节点在 $x=\pm\frac{1}{2}$。注意：它在区间 $\left(-\frac{1}{2},\frac{1}{2}\right)$ 以外恒为 0，见图 1.32 上列。也可以取 0 次反转幂 $x_*^0=\operatorname{sign} x$ 的一阶差分，其结果基本相同：

$$\Delta^1 x_*^0=\operatorname{sign}\left(x+\frac{1}{2}\right)-\operatorname{sign}\left(x-\frac{1}{2}\right)=\begin{cases}(-1)-(-1)=0, & x<-\frac{1}{2}\\ 1-(-1)=2, & -\frac{1}{2}<x<\frac{1}{2}\\ 1-1=0, & \frac{1}{2}<x\end{cases}$$

因此

$$\Delta^1 x_+^0=\frac{1}{2}\Delta^1 x_*^0=\frac{1}{2}\Delta^1 \operatorname{sign} x$$

对 1 次截断幂 x_+^1 作二阶差分，后者定义为

$$\begin{aligned}\Delta^2 f(x)&=\Delta^1(\Delta^1 f(x))=\Delta^1\left(f\left(x+\frac{1}{2}\right)-f\left(x-\frac{1}{2}\right)\right)=\cdots\\&=f(x+1)-2f(x)+f(x-1)\end{aligned}$$

于是

$$\Delta^2 x_+^1=(x+1)_+-2x_++(x-1)_+=\begin{cases}0-2\cdot 0+0=0, & x\leqslant -1\\ x+1-2\cdot 0+0=1+x, & -1\leqslant x\leqslant 0\\ x+1-2x+0=1-x, & 0\leqslant x\leqslant 1\\ x+1-2x+x-1=0, & 1\leqslant x\end{cases}$$

也不难验证

$$\Delta^2 x_+^1=\frac{1}{2}\Delta^2 x_*^1=\frac{1}{2}\Delta^2|x|$$

这是三角形的折线函数(1.3.1 节)，是一次样条，节点在 $x=0,\ \pm1$，它在区间 $(-1,1)$ 以外恒为 0。

对二次截断幂 x_+^2 或 x_*^2 作三阶差分

$$\begin{aligned}\Delta^3 f(x)&=\Delta^1(\Delta^2 f(x))=\Delta^1(f(x+1)-2f(x)+f(x+1))=\cdots\\&=f\left(x+\frac{3}{2}\right)-3f\left(x+\frac{1}{2}\right)+3f\left(x-\frac{1}{2}\right)-f\left(x-\frac{3}{2}\right)\end{aligned}$$

可以得到

$$\Delta^3 x_+^2=\left(x+\frac{3}{2}\right)_+^2-3\left(x+\frac{1}{2}\right)_+^2+3\left(x-\frac{1}{2}\right)_+^2-\left(x-\frac{3}{2}\right)_+^2=\begin{cases}0, & x\leqslant -\frac{3}{2}\\ \left(\frac{3}{2}+x\right)^2, & -\frac{3}{2}\leqslant x\leqslant -\frac{1}{2}\\ \frac{3}{2}-2x^2, & -\frac{1}{2}\leqslant x\leqslant \frac{1}{2}\\ \left(\frac{3}{2}-x\right)^2, & \frac{1}{2}\leqslant x\leqslant \frac{3}{2}\\ 0, & \frac{3}{2}\leqslant x\end{cases}$$

$$\Delta^3 x_+^2=\frac{1}{2}\Delta^3 x_*^2=\frac{1}{2}\Delta^3 x^2\operatorname{sign}x$$

这是二次样条，节点在 $x=\pm\frac{1}{2}$, $\pm\frac{3}{2}$, 它在区间 $\left(-\frac{3}{2},\ \frac{3}{2}\right)$ 以外恒为 0。

对三次截断幂 x_+^3 或 x_*^3 作四阶差分

$$\Delta^4 f(x)=\Delta^1(\Delta^3 f(x))=\Delta^1\left(f\left(x+\frac{3}{2}\right)-3f\left(x+\frac{1}{2}\right)+3f\left(x-\frac{1}{2}\right)-f\left(x-\frac{3}{2}\right)\right)=\cdots$$
$$=f(x+2)-4f(x+1)+6f(x)-4f(x-1)+f(x+2)$$

则有

$$\Delta^4 x_+^3=(x+2)_+^3-4(x+1)_+^3+6x_+^3-4(x-1)_+^3+(x-2)_+^3=\begin{cases}0, & x\leqslant -2\\ (2+x)^3, & -2\leqslant x\leqslant -1\\ (2+x)^3-4(1+x)^3, & -1\leqslant x\leqslant 0\\ (2-x)^3-4(1-x)^3, & 0\leqslant x\leqslant 1\\ (2-x)^3, & 1\leqslant x\leqslant 2\\ 0, & 2\leqslant x\end{cases}$$

$$\Delta^4 x_+^3=\frac{1}{2}\Delta^4 x_*^3=\frac{1}{2}\Delta^4|x|^3$$

这是三次样条，节点在 $x=0$, ±1, ±2, 它在区间 $(-2,\ 2)$ 以外恒为 0。

以上四个样条 $\Delta^1 x_+^0$, $\Delta^2 x_+^1$, $\Delta^3 x_+^2$, $\Delta^4 x_+^3$ 的共同特点是具有紧凑性，即在有界的区间以外恒为 0。此外，它们在不等于 0 之处都取正值，中有高峰，向两方单调递降至 0，成对称的山丘形状，这里，不妨称之为山丘形样条。$\Delta^1 x_+^0$, $\Delta^2 x_+^1$, $\frac{1}{2!}\Delta^3 x_+^2$, $\frac{1}{3!}\Delta^4 x_+^3$ 的曲线形状见图 1.32。

以上结果显然可以推广到一般的 p 次。为此，递推地定义 m 阶差分

$$\Delta^m f(x)=\Delta^1(\Delta^{m-1}f(x)),\ m=2,\ 3,\ \cdots$$

不难验证

$$\Delta^m f(x)=\sum_{k=1}^{m}(-1)^k C_k^m f\left(x+\frac{m}{2}-k\right),\quad C_k^m=\frac{m!}{k!(m-k)!}$$

取 x_+^p 的 $p+1$ 阶差分

$$\Delta^{p+1}x_+^p\equiv\sum_{k=1}^{p+1}(-1)^k C_k^{p+1}\left(x+\frac{p+1}{2}-k\right)_+^p \tag{1.5.6}$$

$p\leqslant4$ 的情况——也是实用上最重要的情况——上面已经有了。这是 p 次山丘形样条，节点在

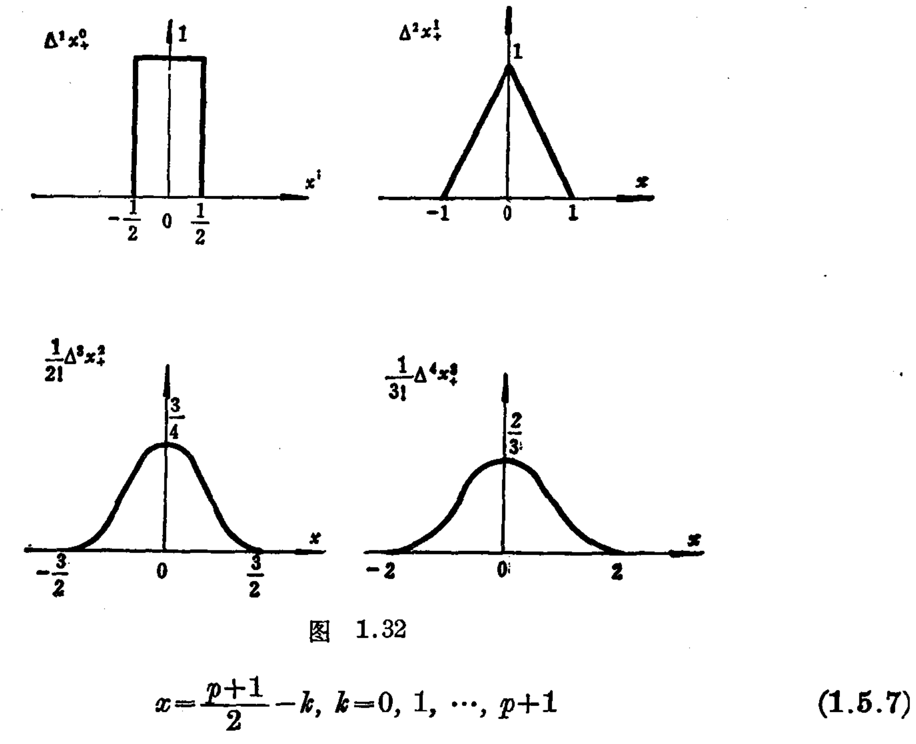


图 1.32

$$x=\frac{p+1}{2}-k,\ k=0,\ 1,\ \cdots,\ p+1 \tag{1.5.7}$$

当 p 为偶数时，节点是“半点”，p 为奇数时，节点为整点，都是等间距的，间距为 1。可以证明

$$\Delta^{p+1}x_+^p\begin{cases}\equiv 0, & \text{当 } |x|\geqslant\dfrac{p+1}{2}\\ >0, & \text{当 } |x|<\dfrac{p+1}{2}\end{cases} \tag{1.5.8}$$

峰点在 $x=0$，对称地向两端单调递降至 0。当 p 增大时，愈来愈光顺，“基底”愈来愈宽。这从图 1.32 也可以看出。

截断幂 x_+^p，或反转幂 x_-^p，或其差分即山丘形样条 $\Delta^{p+1}x_+^p\equiv\frac{1}{2}\Delta^{p+1}x_-^p$，可以作为 p 次样条的基，这就是说，任意 p 次样条都可以表为这些特殊样条及其位移的线性组合。特别重要的是山丘形样条，它们具有紧凑性和计算稳定的优点(1.3.1 节)。以上有关事实的论证及其对于不等间距的推广见 1.5.2 节。关于等间距山丘形样条的性质还可参考第四章 §4.3，那里函数 $\frac{1}{p!}\Delta^{p+1}x_+^p$ 记为 $M_{p+1}(x)$。

1.5.2 分节区间上的一般样条

在 x 轴上取节点

$$t_0<t_1<\cdots<t_{n-1}<t_n \tag{1.5.9}$$

分节区间 $[t_0,\ t_1,\ \cdots,\ t_n]$ 上的 p 次样条，是指在各段 $[t_0,\ t_1]$，…，$[t_{n-1},\ t_n]$ 上分别为次数 $\leqslant p$ 的多项式，它在内节点 t_1，…，t_{n-1} 上直至 $p-1$ 阶导数连续。这种 p 次样条组成一个函数类，记作 $S_p[t_0,\ t_1,\ \cdots,\ t_n]$，简记作 S_p。S_0 就是分段常数的函数类。$[t_0,\ t_1,\ \cdots,\ t_n]$ 有 n 个分段，每个分段上 p 次多项式有 $p+1$ 个系数，共有 $n(p+1)$ 个系数。它们不是互相独立的，在每个内节点要满足 p 个连续条件

$$f(t_k-0)=f(t_k+0),\ f'(t_k-0)=f'(t_k+0),\ \cdots,\ f^{(p-1)}(t_k-0)=f^{(p-1)}(t_k+0)$$
$$k=1,\ 2,\ \cdots,\ n-1 \tag{1.5.10}$$

共计$(n-1)p$个条件,这些条件都是互相独立的,因此S_p内的函数的自由度是

$$n(p+1)-(n-1)p=n+p$$

这就是说,需要$n+p$个条件才能唯一决定$[t_0,\ t_1,\ \cdots,\ t_n]$上的一个$p$次样条。

对$p=0,\ 1,\ 3$的情况,事实上在§1.3中已经讨论过了。

当$p=0$时,需要n个条件,例如取f在半点$t_{i+\frac{1}{2}}=\frac{1}{2}(t_i+t_{i+1})$的值$f_{i+\frac{1}{2}}$, $i=0,\ 1,\ \cdots,\ n-1$。当$p=1$时需要$n+1$个条件,例如取整点t_i的函数值f_i, $i=0,\ 1,\ \cdots,\ n$。当$p=3$时,$n+3$个条件可取为f_i, $i=0,\ 1,\ \cdots,\ n$以及f'_0, f'_n或者f''_0, f''_n。

对于一般的$p=2q+1=$奇数,需要$n+p=n+1+2q$个条件,可以取$f_0,\ f_1,\ \cdots,\ f_n$以及两端各q个"边界条件",如$f'_0,\ \cdots,\ f^{(q)}_0,\ f'_n,\ \cdots,\ f^{(q)}_n$。对于$p=2q=$偶数,需要$n+p=n+2q$个条件。它们可以取,例如$n$个半点值$f_{\frac{1}{2}},\ f_{1+\frac{1}{2}},\ \cdots,\ f_{n-\frac{1}{2}}$,以及两端各$q$个边界条件$f'_0,\cdots,\ f^{(q)}_0,\ f'_n,\ \cdots,\ f^{(q)}_n$;也可以取$n+1$个整点值$f_0,\ \cdots,\ f_n$,以及$f'_0,\ \cdots,\ f^{(q)}_0,\ f'_n,\ \cdots,\ f^{(q-1)}_n$(或$f'_0,\ \cdots,\ f^{(q-1)}_0,\ f'_n,\ \cdots,\ f^{(q)}_n$)。不过,稍为别扭些。

截断幂($i=0,\ 1,\ \cdots,\ n$)

$$(x-t_i)^p_+=\begin{cases}0, & x<t_i\\ (x-t_i)^p, & x>t_i\end{cases} \tag{1.5.11}$$

显然是$[t_0,\ t_1,\ \cdots,\ t_n]$上的$p$次样条,即属于$S_p$,其$p$阶导数在$x=t_i$处间断,跃值为$p!$。注意:$(x-t_n)^p_+$在$[t_0,\ t_n]$上恒为0,不起作用。此外,$1,\ x,\ \cdots,\ x^p$当然也是$p$次样条,属于$S_p$。不难看到,在$[t_0,\ t_1,\ \cdots,\ t_n]$上的任意的$p$次样条$f(x)$可以唯一地表为

$$f(x)=\sum_{i=1}^{n-1}\alpha_i\,(x-t_i)^p_+ +\sum_{k=0}^{p}\beta_k x^k,\quad t_0\leqslant x\leqslant t_n \tag{1.5.12}$$

因此,$n+p$个函数

$$(x-t_1)^p_+,\ \cdots,\ (x-t_{n-1})^p_+,\ 1,\ x,\ \cdots,\ x^p \tag{1.5.13}$$

构成$S_p[t_0,\ t_1,\ \cdots,\ t_n]$的基。

为了证明,任取$f\in S_p$,依次在内节点$x=t_1,\ \cdots,\ t_{n-1}$取其p阶导数的跃值得$[f^{(p)}]_{x=t_1}$, $\cdots,\ [f^{(p)}]_{x=t_{n-1}}$。作函数

$$g(x)=f(x)-\sum_{i=1}^{n-1}\alpha_i(x-t_i)^p_+,\quad \alpha_i=\frac{1}{p!}[f^{(p)}]_{x=t_i},\quad i=1,\ \cdots,\ n-1 \tag{1.5.14}$$

显然,$g(x)$与$f(x)$一样,也是$[t_0,\ t_1,\ \cdots,\ t_n]$上的$p$次样条,即$g(x)$为分段$p$次多项式,而$g,\ g^1,\ \cdots,\ g^{(p-1)}$在分点$t_1,\ \cdots,\ t_{n-1}$连续。试在分点$x=t_1$取$g^{(p)}$的跃值。由于$(x-t_1)^p_+$的$p$阶导数在$x=t_1$处的跃值为$p!$,而其它$(x-t_2)^p_+,\ \cdots,\ (x-t_{n-1})^p_+$的$p$阶导数在$x=t_1$处连续,跃值为0,因此

$$[g^{(p)}]_{x=t_1}=[f^{(p)}]_{x=t_1}-\alpha_1 p!-\alpha_2\cdot 0-\cdots-\alpha_{n-1}\cdot 0=0$$

类似地可知$[g^{(p)}]_{x=t_1}=\cdots=[g^{(p)}]_{x=t_{n-1}}=0$,因此,$g^{(p)}$在分点$t_1,\ \cdots,\ t_{n-1}$连续。由于$g(x)$为分段$p$次多项式并且直到$p$阶导数为连续,因此它在$[t_0,\ t_1,\ \cdots,\ t_n]$上是"真正"的$p$次多项式,故必可表为$g(x)=\sum\limits_{0}^{p}\beta_k x^k$。连系到式(1.5.14)可知,表达式(1.5.12)成立。至于唯一性,假如除式(1.5.12)外$f(x)$还可表为

$$f(x)=\sum_{1}^{n-1}\alpha_i'(x-t_i)_+^p+\sum_{0}^{p}\beta_k' x^k$$

两式相减，

$$0\equiv\sum_{1}^{n-1}(\alpha_i-\alpha_i')(x-t_i)_+^p+\sum_{0}^{p}(\beta_k-\beta_k')x^k,\quad t_0\leqslant x\leqslant t_n$$

依次在分点 $x=t_1, \cdots, t_{n-1}$ 计算上式两端的 p 阶导数的跃值，由与上述相同的推理，并考虑到 x^k 的 p 阶导数恒连续，可以得到 $0=\alpha_1-\alpha_1', \cdots, 0=\alpha_{n-1}-\alpha_{n-1}'$，因此

$$0\equiv\sum_{0}^{p}(\beta_k-\beta_k')x^k, t_0\leqslant x\leqslant t_n$$

因此 $\beta_0-\beta_0'=0, \cdots, \beta_p-\beta_p'=0$，唯一性得证，并且系数 α_i 必为

$$\alpha_i=\frac{1}{p!}[f^{(p)}(t_i+0)-f^{(p)}(t_i-0)], i=1, \cdots, n-1$$

上述 p 次样条的基(1.5.13)含有两种成分 $(x-t_i)_+^p$ 和 x^k，形式上不统一，但可以把它们统一起来。为此，引进"界外"的节点

$$\cdots t_{-2}<t_{-1}<t_0,\ t_n<t_{n+1}<t_{n+2}<\cdots \tag{1.5.15}$$

除了这个条件外，其它为任意，但一次取定。考虑 $n+p$ 个样条

$$(x-t_i)_+^p,\ i=-p,\ -p+1, \cdots, 0, \cdots, n-1 \tag{1.5.16}$$

当 $i=-p, -p+1, \cdots, 0$ 时，$(x-t_i)_+^p$ 在区间 $[t_0, t_1, \cdots, t_n]$ 实质上就是"真正"的多项式 $(x-t_i)^p$。不难证明，它们在 $[t_0, t_1, \cdots, t_n]$ 上是互相独立的。因此多项式 $\sum_{0}^{p}\beta_k x^k$ 可以唯一地表为它们的线性组合。因此式(1.5.16)也构成 $S_p[t_0, t_1, \cdots, t_n]$ 的基。

相同的道理，$S_p[t_0, t_1, \cdots, t_n]$ 的基也可以取"反转幂"的形式，从而得出类似于(1.5.13)的形式

$$(x-t_1)_*^p,\ (x-t_2)_*^p, \cdots, (x-t_{n-1})_*^p, 1, x, \cdots, x^p \tag{1.5.17}$$

或类似于(1.5.16)的形式

$$(x-t_i)_*^p,\ i=-p,\ -p+1, \cdots, 0, \cdots, n-1 \tag{1.5.18}$$

1.5.3 山丘形基样条

在 1.3.1 节中片段零次插值的基函数是矩形函数，片段线性插值的基函数是三角形的。零次样条

$$\pi_{i+\frac{1}{2}}(x)=\begin{cases}0, & x<t_i\\ 1, & t_i<x<t_{i+1}\\ 0, & t_{i+1}<x\end{cases}\quad (i=0, 1, \cdots, n-1) \tag{1.5.19}$$

构成 $S_0[t_0, t_1, \cdots, t_n]$ 的基。一次样条

$$\lambda_i(x)=\begin{cases}0, & x\leqslant t_{i-1}\\ \dfrac{x-t_{i-1}}{t_i-t_{i-1}}, & t_{i-1}\leqslant x\leqslant t_i\\ \dfrac{t_{i+1}-x}{t_{i+1}-t_i}, & t_i\leqslant x\leqslant t_{i+1}\\ 0, & t_{i+1}\leqslant x\end{cases}\quad (i=0, 1, \cdots, n) \tag{1.5.20}$$

构成 $S_1[t_0, t_1, \cdots, t_n]$ 的基。

这两种样条的重要特征在于，它们都是紧凑函数（见 §1.3），即在少数几个节距以外恒为零。因此，数据误差只影响到很局部的范围，基本上不扩散，不放大，具有计算稳定的优点。反之，样条$(x-t_i)_+^p$和$(x-t_i)_*^p$不为零的范围是(t_i, ∞)和$(-\infty, \infty)$，都是无限的，它们都不是紧凑函数，数据误差按p次幂传播到无穷远，有一定的不稳定性，它不利于数值计算。在1.5.1节中，用差分的方法从x_+^p或x_*^p构造了紧凑的山丘形样条。显然可以推广到目前的场合，只需把等距差分改为不等距差商，就可以从$(x-t_i)_+^p$或$(x-t_i)_*^p$构造出紧凑的基样条来，包括作为特例的$\pi_{i+\frac{1}{2}}$和λ_i。

零次样条：作$(x-t_i)_+^0$的反号一阶差商

$$\psi_{0,i}(x)=-\frac{1}{t_{i+1}-t_i}[(x-t_{i+1})_+^0-(x-t_i)_+^0]=\frac{1}{t_{i+1}-t_i}[(x-t_i)_+^0-(x-t_{i+1})_+^0]$$

$$=\beta_{i,i}(x-t_i)_+^0+\beta_{i,i+1}(x-t_{i+1})_+^0 \tag{1.5.21}$$

$$\beta_{i,i}=-\frac{1}{t_i-t_{i+1}}\quad,\quad \beta_{i,i+1}=-\frac{1}{t_{i+1}-t_i}$$

不难验证

$$\psi_{0,i}(x)=\left\{\begin{matrix}0, & x<t_i\\ \dfrac{1}{t_{i+1}-t_i}, & t_i<x<t_{i+1}\\ 0, & t_{i+1}<x\end{matrix}\right\}\equiv\frac{1}{t_{i+1}-t_i}\pi_{i+\frac{1}{2}}(x) \tag{1.5.22}$$

这里的所谓差商，实质上是取两个变量x, t的函数$(x-t)_+^0$，对变量t取节点$t=t_i$, t_{i+1}的差商。取反号只是为了保证结果得正值。

一次样条：作$(x-t_i)_+^1$的二阶差商

$$\psi_{1,i}(x)=\beta_{i,i-1}(x-t_{i-1})_++\beta_{i,i}(x-t_i)_++\beta_{i,i+1}(x-t_{i+1})_+ \tag{1.5.23}$$

$$\beta_{i,i-1}=\frac{2}{(t_{i-1}-t_i)(t_{i-1}-t_{i+1})},\quad \beta_{i,i}=\frac{2}{(t_i-t_{i-1})(t_i-t_{i+1})},\quad \beta_{i,i+1}=\frac{2}{(t_{i+1}-t_{i-1})(t_{i+1}-t_i)}$$

不难验证：

$$\psi_{1,i}(x)=\left\{\begin{matrix}0, & x\leqslant t_{i-1}\\ \dfrac{2(x-t_{i-1})}{(t_{i+1}-t_{i-1})(t_i-t_{i-1})}, & t_{i-1}\leqslant x\leqslant t_i\\ \dfrac{2(x-t_{i+1})}{(t_{i+1}-t_{i-1})(t_i-t_{i+1})}, & t_i\leqslant x\leqslant t_{i+1}\\ 0, & t_{i+1}\leqslant x\end{matrix}\right\}\equiv\frac{2}{t_{i+1}-t_{i-1}}\lambda_i(x) \tag{1.5.24}$$

p次样条：对于一般的$p=0, 1, 2, \cdots$取两个变量$x-t$的函数$(x-t)_+^p$，作对于变量t以$t=t_i, t_{i+1}, \cdots, t_{i+p+1}$为节点的$p+1$阶差商（当$p$为偶数时约定取反号，以保证结果得正值）：

$$\psi_{p,i}(x)=\sum_{j=i}^{i+p+1}\beta_{i,j}(x-t_j)_+^p \tag{1.5.25}$$

此处

$$\left.\begin{aligned}&\beta_{i,j}=(-1)^{p+1}(p+1)!/\omega_i'(t_j)\\&\omega_i(t)=(t-t_i)(t-t_{i+1})\cdots(t-t_{i+p+1})\\&\omega_i'(t_j)=(t_j-t_i)(t_j-t_{i+1})\cdots(t_j-t_{j-1})(t_j-t_{j+1})\cdots(t_j-t_{i+p+1})\end{aligned}\right\} \tag{1.5.26}$$

这就是$p+1$阶数值微商（乘以符号$(-1)^{p+1}$）的系数，见1.5.4节。

显然，对于任意整数 i，$\psi_{p,i}\in S_p[t_0, t_1, \cdots, t_n]$。$\psi_{p,i}$ 的一个重要性质是紧凑性，它在子区间 (t_i, t_{i+p+1}) 以外恒为 0，即

$$\psi_{p,i}(x)\equiv 0,\ \text{当}\ x\leqslant t_i\ \text{或}\ x\geqslant t_{i+p+1} \tag{1.5.27}$$

事实上，根据截断幂的定义(1.5.1)

$$(x-t_j)_+^p=\begin{cases}0, & x\leqslant t_i\\ (x-t_j)^p, & x\geqslant t_{i+p+1}\end{cases}\quad j=i,\ i+1,\ \cdots,\ i+p+1$$

因此，当 $x\leqslant t_i$ 时，$\psi_{p,i}(x)\equiv 0$；而当 $x\geqslant t_{i+p+1}$ 时，$\psi_{p,i}(x)$ 为 t 的 p 次多项式 $(x-t)^p$ 以

$$t=t_i,\ \cdots,\ t_{i+p+1}$$

为节点的 $p+1$ 阶差商(或反号)，也恒等于 0。

此外，还可以证明，在开区间 (t_i, t_{i+p+1}) 上 $\psi_{p,i}$ 恒为正，即

$$\psi_{p,i}(x)>0,\ \text{当}\ t_i<x<t_{i+p+1} \tag{1.5.28}$$

并且有一个唯一的极大点。因此 $\psi_{p,i}$ 是单峰式的山丘形函数(仅当 $p=0$ 时是平顶的)；当 p 增大时，光滑度递增，而"基底"即函数不为零的范围变宽。

也可以取"反转幂"代替"截断幂"来作差商，结果只差一个常数倍(2倍)。事实上，由于式(1.5.25)，(1.5.3)

$$\psi_{p,i}(x)=\frac{1}{2}\sum_{j=i}^{i+p+1}\beta_{i,j}(x-t_j)_*^p+\frac{1}{2}\sum_{j=i}^{i+p+1}\beta_{i,j}(x-t_j)^p$$

右端第二项是 t 的 p 次多项式 $(x-t)^p$ 对 t 的 $p+1$ 阶差商，恒为 0，因此

$$\psi_{p,i}(x)=\frac{1}{2}\sum_{j=i}^{i+p+1}\beta_{i,j}(x-t_j)_*^p=\begin{cases}\dfrac{1}{2}\sum\limits_{j=i}^{i+p+1}\beta_{i,j}|x-t_i|^p\operatorname{sign}(x-t_j), & p=\text{偶数}\\ \dfrac{1}{2}\sum\limits_{j=i}^{i+p+1}\beta_{i,j}|x-t_j|^p, & p=\text{奇数}\end{cases} \tag{1.5.29}$$

当 $p=$ 奇数时用这个公式来计算 $\psi_{p,i}$ 是比较方便的。

根据性质(1.5.27)，(1.5.28)，当 $i\leqslant-(p+1)$ 或 $i\geqslant n$ 时，样条 $\psi_{p,i}$ 在区间 $[t_0, t_n]$ 上恒为 0，不起作用；在 $[t_0, t_n]$ 不恒为 0 的只有下列 $n+p$ 个

$$\psi_{p,i}(x),\ i=-p,\ -p+1,\ \cdots,\ 0,\ 1,\ \cdots,\ n-1 \tag{1.5.30}$$

可以证明(证略)，它们在区间 $[t_0, t_n]$ 上是线性无关的，因此构成 $S_p[t_0, t_1, \cdots, t_n]$ 的基，即任意的 $f(x)\in S_p$ 可以唯一地表为

$$f(x)=\sum_{i=-p}^{n-1}\alpha_i\psi_{p,i}(x),\quad t_0\leqslant x\leqslant t_n \tag{1.5.31}$$

这里基样条都是紧凑的，有利于数值计算。

当 $p=0, 1$ 时是已知的情况(见 1.3.1 节)，系数 α_i 可以简单地通过 $f(x)$ 在半点或整点的值来表达：

$$p=0:\ \alpha_i=(t_{i+1}-t_i)f_{i+\frac{1}{2}},\quad i=0,\ 1,\ \cdots,\ n-1 \tag{1.5.32}$$

$$p=1:\ \alpha_i=\frac{1}{2}(t_{i+2}-t_i)f_{i+1},\ i=-1,\ 0,\ \cdots,\ n-1 \tag{1.5.33}$$

但在 $p\geqslant 2$ 时，则 α_i 没有这样简单的表达式，需要利用定解条件来联解一个线代数方程组，以确定系数 α_i。现举 $p=3$ 为例来加以说明，其它情况是类似的。

对于 $[t_0, t_1, \cdots, t_n]$ 上的三次样条，可以取，比方说，下列 $n+3$ 个定解条件

$$f'(t_0)=f_0',\ f(t_0)=f_0,\ \cdots,\quad f(t_n)=f_n,\ f'(t_n)=f_n' \tag{1.5.34}$$

于是得到线代数方程组

$$\begin{cases}\sum_{i=-3}^{n-1}\alpha_i\psi_i'(t_0)=f_0' \\ \sum_{i=-3}^{n-1}\alpha_i\psi_i(t_k)=f_k \quad k=0,\ 1,\ \cdots,\ n \\ \sum_{i=-3}^{n-1}\alpha_i\psi_i'(t_n)=f_n'\end{cases} \tag{1.5.35}$$

未知数是 $\alpha_{-3},\ \alpha_{-2},\ \cdots,\ \alpha_{n-1}$，系数及右端项都是已知的，这里为了简便，命 $\psi_i=\psi_{3,i}$。

根据性质(1.5.27)以及 p 次样条 $p-1$ 阶以下导数的连续性，可得

$$\psi_{p,i}(t_j)=\psi'_{p,i}(t_j)=\cdots=\psi_{p,i}^{(p-1)}(t_j)=0,\ 当\ j\leqslant i\ 或\ j\geqslant i+p+1 \tag{1.5.36}$$

因此对 $p=3$ 有

$$\psi_i(t_j)=\psi_i'(t_j)=\psi_i''(t_j)=0,当\ j\leqslant i\ 或\ j\geqslant i+4 \tag{1.5.37}$$

于是方程组(1.5.35)中每个方程至多有三个系数不为0，即

$$\begin{cases}\alpha_{-3}\psi'_{-3}(t_0)+\alpha_{-2}\psi'_{-2}(t_0)+\alpha_{-1}\psi'_{-1}(t_0)=f_0' \\ \alpha_{k-3}\psi_{k-3}(t_k)+\alpha_{k-2}\psi(t_k)+\alpha_{k-1}\psi_{k-1}(t_k)=f_k \quad k=0,\ 1,\ \cdots,\ n \\ \alpha_{n-3}\psi'_{n-3}(t_n)+\alpha_{n-2}\psi'_{n-2}(t_n)+\alpha_{n-1}\psi'_{n-1}(t_n)=f_n'\end{cases} \tag{1.5.38}$$

当取定解条件为

$$f''(t_0)=f_0'',\ f(t_0)=f_0,\ \cdots,\ f(t_n)=f_n,\ f''(t_n)=f_n'' \tag{1.5.39}$$

时则得方程组

$$\begin{cases}\alpha_{-3}\psi''_{-3}(t_0)+\alpha_{-2}\psi''_{-2}(t_0)+\alpha_{-1}\psi''_{-1}(t_0)=f_0'' \\ \alpha_{k-3}\psi_{k-3}(t_k)+\alpha_{k-2}\psi_{k-2}(t_k)+\alpha_{k-1}\psi_{k-1}(t_k)=f_k \quad k=0,\ 1,\ \cdots,\ n \\ \alpha_{n-3}\psi''_{n-3}(t_n)+\alpha_{n-2}\psi''_{n-2}(t_n)+\alpha_{n-1}\psi''_{n-1}(t_n)=f_n''\end{cases} \tag{1.5.40}$$

所谓$[t_0,\ t_1,\ \cdots,\ t_n]$上的周期样条，就是以区间$[t_0,\ t_n]$的长度 t_n-t_0 为周期拓至$-\infty<x<\infty$ 上的 p 次样条。这时，节点 $t_0,\ t_1,\ \cdots,\ t_n$ 也应以周期 t_n-t_0 周期地拓至“界外”，应该满足

$$t_{i+n}=t_i+(t_n-t_0) \tag{1.5.41}$$

由于幂样条 $\varphi_i(x)=(x-t_i)_+^p$ 具有平移不变性

$$\varphi_{i+n}(x)=\varphi_i(x-(t_n-t_0))$$

并且它们的差商基样条 $\psi_i(x)$ 也具有平移不变性

$$\psi_{i+n}(x)=\psi_i(x-(t_n-t_0))$$

因此

$$\psi_{i+n}(t_j)=\psi_i(t_{j-n}) \tag{1.5.42}$$

这样，$[t_0,\ t_1,\ \cdots,\ t_n]$上的周期3次样条的系数 $\alpha_i(i=-3,\ \cdots,\ n-1)$ 虽然还是 $n+3$ 个，但满足周期性条件

$$\alpha_{i+n}=\alpha_i \tag{1.5.43}$$

故只有 n 个是独立的，例如取为 $\alpha_{-1},\ \alpha_0,\ \cdots,\ \alpha_{n-2}$。定解条件也只需 n 个，例如取为

$$f(t_1)=f_1,\ f(t_2)=f_2,\ \cdots,\ f(t_n)=f_n \tag{1.5.44}$$

于是 $\alpha_{-1},\ \cdots,\ \alpha_{n-2}$ 满足方程组

$$\begin{cases}\alpha_{n-2}\psi_{-2}(t_1)+\alpha_{-1}\psi_{-1}(t_1)+\alpha_0\psi_0(t_1)=f_1\\ \alpha_{k-3}\psi_{k-3}(t_k)+\alpha_{k-2}\psi_{k-2}(t_k)+\alpha_{k-1}\psi_{k-1}(t_k)=f_k,\ k=2,\ \cdots,\ n-1\\ \alpha_{n-3}\psi_{n-3}(t_n)+\alpha_{n-2}\psi_{n-2}(t_n)+\alpha_{-1}\psi_{n-1}(t_n)=f_n\end{cases} \tag{1.5.45}$$

而

$$\alpha_{-3}=\alpha_{n-3},\ \alpha_{-2}=\alpha_{n-2},\ \alpha_{n-1}=\alpha_{-1} \tag{1.5.46}$$

以上三种情况的方程组基本上是三对角线带状的。当节点距离没有什么突变时，系数阵的对角线元占优势，因此是比较容易解的（比较1.3.2节以及以下等距的情况）。

1.5.4 等间距的基样条

当节点为等距 $t_{i+1}-t_i=h$ 时，不难验证系数 β_{ij} 就是二项系数

$$\beta_{i,i+k}=\frac{(-1)^{2(p+1)-k}}{h^{p+1}}C_k^{p+1},\qquad C_k^{p+1}=\frac{(p+1)!}{k!(p+1-k)!} \tag{1.5.47}$$

这时，p 同而 i 不同的 $\psi_{p,i}$ 之间，彼此只差一个平移，即

$$\psi_{p,i}(x)\equiv\psi_{p,0}(x-ih) \tag{1.5.48}$$

$\psi_{p,i}(x)$ 对称于区间 $[t_i,\ t_{i+p+1}]$ 的中点 $\frac{1}{2}(t_i+t_{i+p+1})=t_i+\frac{(p+1)h}{2}$，当 $p=$ 偶数时这是半点，当 $p=$ 奇数时这是整点。当 $p=0,\ 1,\ 2,\ 3$ 时 $\psi_{p,i}$ 的表达式为

$$p=0:\ \psi_{0,i}=\frac{1}{h}[(x-t_i)_+^0-(x-t_{i+1})_+^0] \tag{1.5.49}$$

$$p=1:\ \psi_{1,i}=\frac{1}{h^2}[(x-t_i)_+-2(x-t_{i+1})_++(x-t_{i+2})_+] \tag{1.5.50}$$

$$p=2:\ \psi_{2,i}=\frac{1}{h^3}[(x-t_i)_+^2-3(x-t_{i+1})_+^2+3(x-t_{i+2})_+^2-(x-t_{i+3})_+^2] \tag{1.5.51}$$

$$p=3:\ \psi_{3,i}=\frac{1}{h^4}[(x-t_i)_+^3-4(x-t_{i+1})_+^3+6(x-t_{i+2})_+^3-4(x-t_{i+3})_+^3+(x-t_{i+4})_+^3] \tag{1.5.52}$$

其形状大致如图1.32，只是比例尺度的不同。

对于 $p=3$，命 $\psi_{3,i}=\psi_i$

$$\psi_i(t_i)=0,\quad \psi_i(t_{i+1})=\frac{1}{h},\quad \psi_i(t_{i+2})=\frac{4}{h},\quad \psi_i(t_{i+3})=\frac{1}{h},\quad \psi_i(t_{i+4})=0$$

$$\psi_i'(t_i)=0,\quad \psi_i'(t_{i+1})=\frac{3}{h^2},\quad \psi_i'(t_{i+2})=0,\quad \psi_i'(t_{i+3})=-\frac{3}{h^2},\quad \psi_i'(t_{i+4})=0$$

$$\psi_i''(t_i)=0,\quad \psi_i''(t_{i+1})=\frac{6}{h^3},\quad \psi_i''(t_{i+2})=-\frac{12}{h^3},\quad \psi_i''(t_{i+3})=\frac{6}{h^3},\quad \psi_i''(t_{i+4})=0$$

取(1.5.34)为定解条件时方程组(1.5.38)成为

$$\begin{bmatrix}-3&0&3&&&\\1&4&1&&&\\&1&4&1&&\\&&\ddots&\ddots&\ddots&\\&&&1&4&1\\&&&-3&0&3\end{bmatrix}\begin{bmatrix}\alpha_{-3}\\\alpha_{-2}\\\alpha_{-1}\\\vdots\\\alpha_{n-2}\\\alpha_{n-1}\end{bmatrix}=\begin{bmatrix}h^2f_0'\\hf_0\\hf_1\\\vdots\\hf_n\\h^2f_n'\end{bmatrix} \tag{1.5.53}$$

将头两个和末两个方程分别相消，得等价的方程组

$$\begin{bmatrix} 6 & 12 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & & 12 & 6 \end{bmatrix} \begin{bmatrix} \alpha_{-3} \\ \alpha_{-2} \\ \alpha_{-1} \\ \vdots \\ \alpha_{n-2} \\ \alpha_{n-1} \end{bmatrix} = \begin{bmatrix} h(3f_0 - hf'_0) \\ hf_0 \\ hf_1 \\ \vdots \\ hf_n \\ h(3f_n + hf'_n) \end{bmatrix} \tag{1.5.54}$$

取(1.5.39)为定解条件时方程组(1.5.40)成为

$$\begin{bmatrix} 6 & -12 & 6 & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & 6 & -12 & 6 \end{bmatrix} \begin{bmatrix} \alpha_{-3} \\ \alpha_{-2} \\ \alpha_{-1} \\ \vdots \\ \alpha_{n-2} \\ \alpha_{n-1} \end{bmatrix} = \begin{bmatrix} h^3 f''_0 \\ hf_0 \\ hf_1 \\ \vdots \\ hf_n \\ h^3 f''_n \end{bmatrix} \tag{1.5.55}$$

等价于

$$\begin{bmatrix} 0 & 36 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & & 36 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{-3} \\ \alpha_{-2} \\ \alpha_{-1} \\ \vdots \\ \alpha_{n-2} \\ \alpha_{n-1} \end{bmatrix} = \begin{bmatrix} h(6f_0 - h^2 f''_0) \\ hf_0 \\ hf_1 \\ \vdots \\ hf_n \\ h(6f_n - h^2 f''_n) \end{bmatrix}$$

将 α_{-3} 与 α_{-2}, α_{n-2} 与 α_{n-1} 对调得

$$\begin{bmatrix} 36 & 0 & & & & & \\ 4 & 1 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 & \\ & & & & 1 & 1 & 4 \\ & & & & & 0 & 36 \end{bmatrix} \begin{bmatrix} \alpha_{-2} \\ \alpha_{-3} \\ \alpha_{-1} \\ \vdots \\ \alpha_{n-3} \\ \alpha_{n-1} \\ \alpha_{n-2} \end{bmatrix} = \begin{bmatrix} h(6f_0 - h^2 f''_0) \\ hf_0 \\ hf_1 \\ \vdots \\ hf_{n-1} \\ hf_n \\ h(6f_n - h^2 f''_n) \end{bmatrix} \tag{1.5.56}$$

对周期样条，条件(1.5.43)与方程(1.5.45)成为

$$\begin{bmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{bmatrix} \begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \vdots \\ \alpha_{n-3} \\ \alpha_{n-2} \end{bmatrix} = \begin{bmatrix} hf_1 \\ hf_2 \\ \vdots \\ hf_{n-1} \\ hf_n \end{bmatrix} \tag{1.5.57}$$

§1.6 多项式和样条的最小二乘法

1.6.1 最小二乘问题

设有一组节点 $x_1<x_2<\cdots<x_m$ 以及相应的实验值 $f_1, f_2, \cdots, f_m$。把样点 (x_1, f_1), (x_2, f_2), $\cdots$, (x_m, f_m) 视为 x-y 平面上的一组点，要求用一条比较简单的曲线 $y=F(x)$，例如多项式曲线去逼近这组点。当数据量 m 较大时，要求作“过点”的，即通过所有样点的多项式插值是不现实的，因为多项式插值在高次时不稳定；同时，这种“过点”的情况也是不可取的，因为实验数据总含有误差(不妨称之为噪音)，所以无须要求完全密合，相反地还需要进行平滑以滤去噪音。因此，通常希望作低次的多项式，它的次数常远小于 m，使得它在一定的意义下最优地逼近于原始数据。

逼近的度量可以有种种取法。最重要同时也是最简单的一种是取所谓方差，即各点偏差的平方和

$$\sum_{k=1}^{m}[F(x_k)-f_k]^2 \tag{1.6.1}$$

有时，由于各点数据的可靠性不一致，故可以权衡轻重而引进权数，即比重 $d_1, d_2, \cdots, d_m>0$。对于较可靠的点赋以较大的比重，从而引用加了权的方差

$$\sum_{k=1}^{m}d_k[F(x_k)-f_k]^2 \tag{1.6.2}$$

作为逼近的度量。

所谓最小二乘或最小平差问题是指 $F(x)$ 取在某特定的函数类中，要求定出一个使得方差(1.6.2)达到极小的函数。一定次数(例如 $n-1$ 次)的多项式

$$F(x)=\sum_{i=0}^{n-1}a_ix^i \tag{1.6.3}$$

的最小二乘问题就是要定出系数 $a_0, a_1, \cdots, a_{n-1}$ 使方差(1.6.2)为极小。一般地说，选定 n 个函数

$$\varphi_1(x), \cdots, \varphi_n(x) \tag{1.6.4}$$

叫做基函数，考虑它们的线性组合

$$F(x)=\sum_{j=1}^{n}u_j\varphi_j(x) \tag{1.6.5}$$

要求定出其中之一，即定出系数 $u_1, \cdots, u_n$ 使方差(1.6.2)达到极小。这时命

$$a_{kj}=\varphi_j(x_k), \quad k=1, \cdots, m; \quad j=1, \cdots, n \tag{1.6.6}$$

连同 $f_1, \cdots, f_m, d_1, \cdots, d_m$ 都是已知量。于是方差(1.6.2)显然是 $u_1, \cdots, u_n$ 的二次函数，命之为

$$\begin{aligned}J=J(u_1, \cdots, u_n)&=\sum_{k=1}^{m}d_k[F(x_k)-f_k]^2=\sum_{k=1}^{m}d_k[\sum_{j=1}^{n}u_j\varphi_j(x_k)-f_k]^2\\&=\sum_{k=1}^{m}d_k[\sum_{j=1}^{n}a_{kj}u_j-f_k]^2\end{aligned} \tag{1.6.7}$$

根据微积分中的极值原理，使 J 达到极小的 $u_1, \cdots, u_n$ 必满足下列方程组

$$\frac{\partial J}{\partial u_i}=0, \quad i=1, \cdots, n \tag{1.6.8}$$

不难算出

$$\frac{\partial J}{\partial u_i}=2\sum_{k=1}^{m}d_k\left(\sum_{j=1}^{n}a_{kj}u_j-f_k\right)a_{ki}=2\left[\sum_{j=1}^{n}\left(\sum_{k=1}^{m}a_{ki}d_ka_{kj}\right)u_j-\sum_{k=1}^{m}a_{ki}d_kf_k\right]$$

因此，命

$$c_{ij}=\sum_{k=1}^{m}a_{ki}d_ka_{kj},\quad i,\ j=1,\ \cdots,\ n \tag{1.6.9}$$

$$g_i=\sum_{k=1}^{m}a_{ki}d_kf_k,\qquad i=1,\ \cdots,\ n \tag{1.6.10}$$

则 $u_1,\ \cdots,\ u_n$ 满足下列线代数方程组，叫做法方程组

$$\sum_{j=1}^{n}c_{ij}u_j=g_i,\quad i=1,\ \cdots,\ n \tag{1.6.11}$$

引用记号 c_{ij}, g_i, 方差函数(1.6.7)可以表为

$$J=J(u_1,\ \cdots,\ u_n)=\sum_{i=1}^{n}u_i\left(\sum_{j=1}^{n}c_{ij}u_j\right)-2\sum_{i=1}^{n}u_ig_i+\sum_{k=1}^{m}d_kf_k^2 \tag{1.6.12}$$

因此，当 $u_1,\ \cdots,\ u_n$ 为问题的解即满足(1.6.11)时所达到的方差最小值是

$$J_{\min}=-\sum_{i=1}^{n}u_ig_i+\sum_{k=1}^{m}d_kf_k^2 \tag{1.6.13}$$

法矩阵 $C=[c_{ij}]$ 是 n 阶对称方阵。可以证明，当基函数 $\varphi_1(x)$, $\cdots$, $\varphi_n(x)$ 在节点集合 $\{x_1,\ \cdots,\ x_m\}$ 上具有一定的独立性时，法矩阵 C 是正定的。这样，最小二乘问题归结于求解法方程组(1.6.11)的问题，数值解法见第八章。

用矩阵记号时，式(1.6.9～1.6.13)可以表为

$$C=A^TDA \tag{1.6.14}$$

$$g=A^TDf \tag{1.6.15}$$

$$Cu=g \tag{1.6.16}$$

$$J=J(u)=u^TCu-2u^Tg+f^TDf \tag{1.6.17}$$

$$J_{\min}=-u^Tg+f^TDf \tag{1.6.18}$$

这里 $u=[u_i]$, $g=[g_i]$ 是 n 阶列阵，$f=[f_i]$ 是 m 阶列阵，$A=[a_{ij}]$ 是 m 行 n 列长方阵，D 是以 $d_1,\ \cdots,\ d_m$ 为对角元的对角阵，$C=[c_{ij}]$ 是 n 阶对称方阵。上标 T 表示矩阵的转置。

1.6.2 多项式的最小二乘法

考虑多项式的情况。这相当于取

$$\varphi_j(x)=x^{j-1},\quad j=1,\ \cdots,\ n \tag{1.6.19}$$

$$a_{kj}=\varphi_j(x_k)=x_k^{j-1},\quad k=1,\ \cdots,\ m;\quad j=1,\ \cdots,\ n \tag{1.6.20}$$

连同式(1.6.11)，表面看来这是一个比较简单的计算问题。实践表明，事情远非如此。事实上，当 $n\leqslant 4$ 或 5 时情况是正常的，但当 $n\geqslant 6$ 或 7 时开始出现反常，法方程组很“难解”，有效数位的丢失很严重。

为了分析这一情况，取 $x_1,\ x_2,\ \cdots,\ x_m$ 为区间[0, 1]上的等距节点

$$x_1=0,\ \cdots,\quad x_k=\frac{k-1}{m-1},\ \cdots,\quad x_m=1 \tag{1.6.21}$$

并取 $d_1=\cdots=d_m=1$，于是

$$c_{ij}=\sum_{k=1}^{m}a_{ki}a_{kj}=\sum_{k=1}^{m}x_k^{i-1}x_k^{j-1}=\sum_{k=1}^{m}x_k^{i+j-2}$$

$$=(m-1)\frac{1}{(m-1)}\sum_{k=1}^{m}x_k^{i+j-2}\approx(m-1)\cdot\frac{1}{i+j-1}\qquad i,\ j=1,\ \cdots,\ n$$

于是当 $m\to\infty$ 时渐近地 $C=C_n\approx(m-1)\boldsymbol{H}_n$

$$H_n=\begin{bmatrix}1 & \frac{1}{2} & \cdots & \frac{1}{n}\\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1}\\ \vdots & & & \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1}\end{bmatrix}\tag{1.6.22}$$

通常叫做 Hilbert 阵。它是对称正定的，并且知道它的逆矩阵 $\boldsymbol{H}_n^{-1}$ 的明显表达式。$\boldsymbol{H}_n^{-1}$ 的模量最大与最小元素的比值是

$$\left[\left(\frac{3n-1}{2}\right)!\right]^2\ :\ n^3\left[\left(\frac{n-1}{2}\right)!\right]^6$$

对于 $7<n<15$，这个比值的量级≈10^n。当方程组(1.6.11)的右项某个 g_i 有误差 1，就会导致解的误差达 10^n，有巨量的放大！笼统地说，字长为十进制 n 位的机器上能解的阶数限度是 n。显然这并不是由具体解算方法带来的，而是由于方程的解对于右项的扰动异常敏感。

通常，对于这一类计算问题，当解题所要费的劲远超过从问题的简单外表所能设想的程度时，称之为病态的。上述法方程组的问题可以说是病态的。

对称正定矩阵的最大与最小本征值的比值 p 叫做状态数或条件数，可以作为矩阵病态程度的衡量。p 值愈大，病态愈甚。对于 $\boldsymbol{H}_n$ 可以证明

$$p(\boldsymbol{H}_n)\approx e^{3.5n}\tag{1.6.23}$$

即状态数随 n 作指数状增长。这是典型的病态矩阵(参看第八章)。

上述法方程组的病态问题可以通过更换基函数而得到改善。仍以等距节点(1.6.21)为例。我们知道，任意的 $n-1$ 次多项式 $F(x)$ 可以唯一地表为

$$F(x)=\sum_{i=0}^{n-1}\beta_i p_i(x)\tag{1.6.24}$$

这里 $p_i(x)$ 是区间[0, 1]上的 i 次正交多项式，即勒让德多项式。它满足

$$\int_0^1 p_i(x)p_j(x)dx=\delta_{ij}=\begin{cases}0, & 当\ i\neq j\\ 1, & 当\ i=j\end{cases}\tag{1.6.25}$$

(见第二章或参考资料[5])。因此，作 $n-1$ 次多项式的最小二乘法时，可取基函数

$$\varphi_i(x)=p_{i-1}(x),\qquad i=1,\ 2,\ \cdots,\ n\tag{1.6.26}$$

以代替(1.6.19)。这时

$$a_{kj}=\varphi_j(x_k)=p_{j-1}(x_k),\qquad k=1,\ \cdots,\ m;\quad j=1,\ \cdots,\ n\tag{1.6.27}$$

于是

$$c_{ij}=\sum_{k=1}^{m}p_{i-1}(x_k)p_{j-1}(x_k)\approx(m-1)\int_0^1 p_{i-1}(x)p_{j-1}(x)dx=(m-1)\delta_{ij}$$

即 $C=C_n\approx(m-1)I_n$，I_n 为 n 阶单位阵。因此法方程组是良态的。这个例子说明了，在同

一函数类中更换基函数，法方程组的稳定性可有很大变化。

应该指出，上述最小二乘问题也可以看作矛盾线性代数方程组

$$\sum_{j=1}^{n} a_{ij}u_j = f_i, \quad i=1, \cdots, m, \quad a_{ij}=x_i^j$$

在最小二乘的意义下定解的问题。当 $m>n$ 时，方程组一般无解，因此叫做矛盾方程组。但是，可以要求定出 $u_1, \cdots, u_n$ 使得误差平方和或加权误差平方和达到极小

$$J(u_1, \cdots, u_n) = \sum_{i=1}^{m} d_i \left(\sum_{j=1}^{n} a_{ij}u_j - f_i\right)^2 = \text{极小}$$

前面列出的法方程组只是问题的一种等价形式而并非解题必经之路。事实上，对于这类问题存在不经由法方程组的，直接与原始的长方阵 $\boldsymbol{A}=[a_{ij}]$ 打交道的比较稳定的解法(详见第八章)。对于多项式最小二乘问题，宜于采用那里所指出的数值解法。

一般说来，对于多项式平差问题，和对于多项式插值问题一样，不宜盲目地追求高次。这是因为，高次多项式是解析函数，它在极小范围内的性质足以决定其全局的行为，因此，总是比较"别扭"的，它缺乏灵活适应的能力。当次数提高时，虽然在控制点即样点上压低了平差，但在这些控制点之间或其外则往往出现不合理的波动扭拐。以下将介绍另一种途径，即分段多项式的途径，它可以基本上克服上述困难。

1.6.3 样条的最小二乘法

在有些问题中，数据列 $(x_1, f_1), \cdots, (x_m, f_m)$ 反映一个比较"长"(时间上或空间上)的过程。除了实验误差外，本身含有许多波动起伏转折。这种特点是不能随着噪音一起被滤掉，而是要保留和复原的。这就是多自由度的数据拟合问题。它要求待定参数的个数 n 比较大(虽然仍有 $n<m$)。如上所述，采用提高多项式次数的办法是不合适的。但是分段多项式，包括样条在内，具有较好的灵活性和稳定性，在本质上是适应于这一类问题的。特别是三次样条，具备了一般说来足够的光滑度，而且比较简单，对于许多拟合问题是适用的。以下主要谈三次样条的平差，推广到一般 p 次样条也是不难的。

适当地选取分段节点

$$\cdots < t_{-3} < t_{-2} < t_{-1} < t_0 < \cdots < t_r < t_{r+1} < \cdots$$

考虑分节区间 $[t_0, t_1, \cdots, t_r]$ 上的三次样条函数类 $S_3[t_0, t_1, \cdots, t_r]$。应安排这些分段节点 t_i 使得全部数据节点 $x_1, \cdots, x_m$ 都落在区间 $[t_0, t_r]$ 之上，并使 x_1 重合或接近于 t_0，x_m 重合或接近于 t_r。应注意区间分两类不同使命的节点即原始的数据节点 x_j 和样条分段节点 t_i，两者可以不相重。前者是问题中给定的，后者可由解题人主动掌握。我们将在样条类 $S_3[t_0, t_1, \cdots, t_r]$ 上作最小平差，它有自由度 $r+3=n$。S_3 有多种形式的基，如(1.5.13)，(1.5.16)，(1.5.30)等等。从计算稳定性考虑，采用山丘形紧凑基样条(1.5.30)

$$\psi_{3,i}(x) = \sum_{j=i}^{i+4} \beta_{i,j}(x-t_j)_+^3 = \frac{1}{2}\sum_{j=i}^{i+4} \beta_{i,j}|x-t_j|^3, \quad i=-3, -2, \cdots, r-1 \tag{1.6.28}$$

$$\beta_{i,j} = \frac{4!}{\omega_i'(t_j)}, \quad \omega_i'(t_j) = \prod_{\substack{k=i \\ k\neq j}}^{i+4} (t_j - t_k) \tag{1.6.29}$$

即取

$$\varphi_j(x) = \psi_{3,j-4}(x), \quad j=1, 2, \cdots, n \tag{1.6.30}$$

作为最小平差的基函数。

$$F(x)=\sum_{j=1}^{n}u_j\varphi_j(x)=\sum_{j=1}^{n}u_j\psi_{3,j-4}(x),\ t_0\leqslant x\leqslant t_r=t_{n-3} \tag{1.6.31}$$

$$a_{kj}=\varphi_j(x_k)=\psi_{3,j-4}(x_k),\quad k=1,\ \cdots,\ m;\quad j=1,\ \cdots,\ n \tag{1.6.32}$$

以下就按标准程式(1.6.9～1.6.11)进行。

由于$\psi_{3,i}$在区间$(t_i,\ t_{i+4})$外恒为0,所以

$$\psi_{3,i}(x)\cdot\psi_{3,j}(x)\equiv 0,\ 当\ |i-j|>3 \tag{1.6.33}$$

$$\varphi_i(x)\varphi_j(x)\equiv 0,\qquad 当\ |i-j|>3 \tag{1.6.34}$$

这就使a_{ij}, c_{ij}, g_i的计算简化。

为了分析法矩阵$C=C_n$的性状，不妨取$[t_0,\ t_r]=[0,\ 1]$, x_k取为$[0,\ 1]$的等距点如(1.6.21), $d_k\equiv 1$。于是有

$$c_{ij}=\sum_{k=1}^{m}\varphi_i(x_k)\varphi_j(x_k)\approx(m-1)\int_0^1\varphi_i(x)\varphi_j(x)dx=(m-1)b_{ij}$$

因此当m相当大时$C_n\approx(m-1)B_n$, $B_n=[b_{ij}]$。

当$i=j$时两个山丘φ_i, φ_j相重合,这时积分值即B_n的对角元

$$b_{ii}=\int_0^1\varphi_i\varphi_i dx$$

比较大。当$|i-j|$从0起逐渐增大时,φ_i与φ_j逐渐"错开",积分值

$$b_{ij}=\int_0^1\varphi_i\varphi_j dx$$

逐步减小,直至$|i-j|>3$后$\varphi_i\varphi_j\equiv 0$, 积分值$b_{ij}=0$。因此$B$是3对角线的带状阵,对角线元占优势。这个结论对于法矩阵C本身也是成立的。不难证明,B_n和C_n都是对称正定的。此外,可以证明对于一切n, B_n的最大本征值有上界,最小本征值有正的下界,因此B_n的状态数p是有界的,即

$$p(B_n)=O(1) \tag{1.6.35}$$

因此,B_n以及C_n是良态的,便于解算。

顺便指出,如果采用"截断幂"如$(x-t_j)_+^3$或$|x-t_j|^3$作为平差的基函数,则相应的法矩阵C_n和B_n就不再呈带状,也没有对角元的优势;并且可以证明,当$n\to\infty$时B_n的状态数p按n的幂次增长

$$p(B_n)\approx O(n^8) \tag{1.6.36}$$

因此它的病态程度虽比多项式时的H_n(1.6.23)好些,但远劣于取山丘形样条(1.6.35)为基的情况。

最后说明一下在等间距$t_{i+1}-t_i\equiv h$时计算方面的简化。这时基函数$\varphi_j=\psi_{3,j-4}$可以从同一个函数用平移的方法产生。因此先定义一个标准的函数

$$\Psi(x)=\begin{cases}0,\ 当|x|\geqslant 2h\\ |x+2h|^3-4|x+h|^3+6|x|^3-4|x-h|^3+|x-2h|^3,\ 当|x|<2h\end{cases} \tag{1.6.37}$$

这是峰点在$x=0$, 间距为$4h$的三次对称样条。所有的基函数(略去一个非本质的公因子)可由$\Psi(x)$用平移产生。由于$\varphi_j=\psi_{3,j-4}$的峰点在$t_{j-4+2}=t_{j-2}$因此

$$\varphi_j(x)=\Psi(x-t_{j-2}),\quad j=1,\ 2,\ \cdots,\ n \tag{1.6.38}$$

$$a_{kj}=\varphi_j(x_k)=\Psi(x_k-t_{j-2}),\quad k=1,\ \cdots,\ m;\quad j=1,\ \cdots,\ n \tag{1.6.39}$$

$$c_{ij}=\begin{cases}0, & \text{当 } |i-j|>3 \\ \sum_{k=1}^{m} d_k a_{ki} a_{kj}, & \text{当 } |i-j|\leqslant 3\end{cases} \tag{1.6.40}$$

注意: 在(1.6.37)和(1.6.40)中各自的第一式在理论上被其第二式所保证，因此是多余的；但在程序中仍宜加上第一式那样的判断，以避免误差和减少工作量。

参 考 资 料

[1] Зельдович-Мышкс, "Элементы прикладной математики", 1965.
[2] Stiefel, "Introduction to Numerical Mathematics", 1963.
[3] Ahlberg-Nielson-Walsh, "Theory of Splines and Its Applications", 1967.
[4] 北京大学、清华大学编，《计算方法》，科学出版社，1973。

第二章　数值积分

§2.1 引　　言

定积分

$$I=\int_a^b f(x)dx \tag{2.1.1}$$

的计算，从一般的数学基础书籍中，我们知道，通常是由确定被积函数$f(x)$的原函数，即确定一个具有导数$F'(x)=f(x)$的函数$F(x)$，然后计算$F(x)$在$x=a$和$x=b$上的值来得到它的量值：$\int_a^b f(x)dx=F(b)-F(a)$。但在实际计算问题中，这样的做法往往是行不通的。因为在大多数的问题中，被积函数$f(x)$的原函数往往不能由明显的表达式给出，有些即使可以用明显的式子表达，但其表达式过于繁复而不适于计算。甚至在有些问题中，被积函数$f(x)$是不明显知道的：它或者是由数表给出，即只知道它在若干离散点上的量值，或者是定义为某个微分方程的解(而这个微分方程是不能明显解出的)。对于这样的问题，自然更无法应用上述方法了。因此在实际计算中，我们就常常采用另外一种途径——数值积分法。

数值积分法，是从近似计算的角度，采用某种数值过程来求出定积分的近似值。

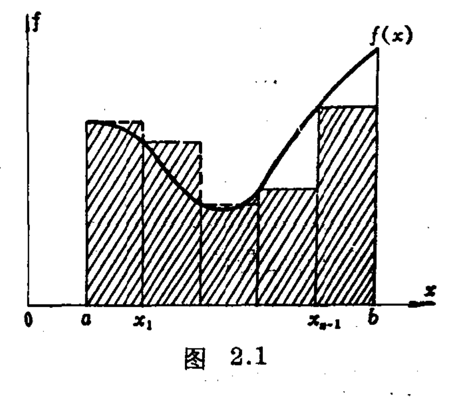


图 2.1

我们知道，定积分$\int_a^b f(x)dx$，当积分区间(a, b)是有限区间，被积函数$f(x)$是(a, b)上的有界函数时，在几何上可以解释为在$x=a$和$x=b$之间函数$f(x)$图形下的面积。如图2.1所示，我们用一串分点

$$a=x_0, x_1, x_2, \cdots, x_n=b$$

将区间(a, b)分成n个小区间：$[x_0, x_1]$，$[x_1, x_2]$，$\cdots$，$[x_{n-1}, x_n]$，并且在各小区间$[x_{i-1}, x_i]$上任取一点ξ_i，设相应的函数值为$f(\xi_i)$，于是得到一组以$h_i=x_i-x_{i-1}$为底，以$f(\xi_i)$为高的矩形，其面积为$h_i f(\xi_i)$。然后，我们将这些矩形面积相加起来，便得到和式

$$I_n=\sum_{i=1}^n h_i f(\xi_i) \tag{2.1.2}$$

设h是n个小区间中的最大长度$h=\max\limits_i\{h_i\}$。如果对于区间的任何一种分法和ξ_i的任何一种选择，当分点无限增多且$h\to 0$时，存在共同极限

$$\lim_{n\to\infty} I_n=\lim_{n\to\infty}\sum_{i=1}^n h_i f(\xi_i)=I$$

则此极限即是定积分$\int_a^b f(x)dx$，且称$f(x)$在(a, b)上是黎曼可积的。

一个有界函数为黎曼可积的充要条件是$f(x)$几乎处处连续。显然，连续函数是黎曼可积的。

从定积分的定义中，我们显然可以得到一种近似计算方法。例如，我们可取分点 x_i 为区间(a, b)的 n 个等分点：$x_i=a+ih\left(h=\frac{b-a}{n}\right)$，并取 ξ_i 为各小区间的两端点之一，或者取 ξ_i 为小区间的中点，那末我们就得到矩形积分公式(或称原始积分公式)：

$$I_n=h\sum_{i=0}^{n-1}f(a+ih) \tag{2.1.3}$$

$$I_n=h\sum_{i=1}^{n}f(a+ih) \tag{2.1.4}$$

$$I_n=h\sum_{i=1}^{n}f\left(a+\left(i+\frac{1}{2}\right)h\right) \tag{2.1.5}$$

其中公式(2.1.5)又称中点积分公式。

直观地，如果我们不用矩形而改用如图 2.2 所示的梯形，那末我们就可以得到定积分的一个较好的近似——梯形积分公式

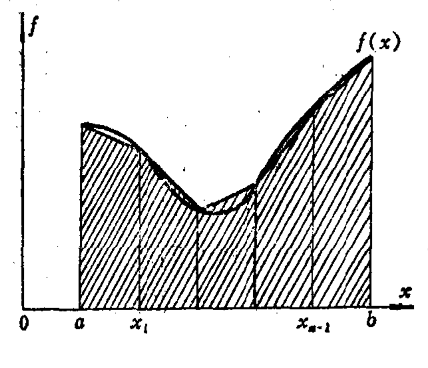


图 2.2

$$I_n=\frac{h}{2}\sum_{i=1}^{n}[f(x_{i-1})+f(x_i)]$$

$$=h\left[\frac{1}{2}f(a)+f(x_1)+f(x_2)+\cdots+f(x_{n-1})+\frac{1}{2}f(b)\right]$$

如此等等。总起来说，各种各样的数值积分法都是利用被积函数值的一种线性组合来近似积分，即

$$I_n=\sum_{i=1}^{n}W_i f(x_i)\approx\int_a^b f(x)dx \tag{2.1.6}$$

此处 $x_1, x_2, \cdots, x_n$ 称为积分坐标点，或称结点；$W_1, W_2, \cdots, W_n$ 称为积分系数，或称伴随于这些结点的“权”。

如上所述，数值积分法是用一有限项的求和计算来代替积分计算，这之间就存在一定的误差。这误差来源于如下两个方面：

第一方面是产生于用有限项之和来等于积分的近似中，这时

$$\int_a^b f(x)dx=I_n+E=\sum_{i=1}^{n}W_i f(x_i)+E \tag{2.1.7}$$

这个误差 E 称为截断误差。

第二方面是产生于我们在计算和式 $I_n=\sum_{i=1}^{n}W_i f(x_i)$ 时是近似地计算，而非精确计算。这是因为计算机是有限字长的，在求和与函数值的计算中带来了舍入误差。因此我们实际得到的是 I_n^*。这时，

$$I_n=I_n^*+R=\sum_{i=1}^{n}{}^*W_i f^*(x_i)+R \tag{2.1.8}$$

这个误差 R 称为舍入误差。

因此，总的误差是

$$\int_a^b f(x)dx-I_n^*=E+R \tag{2.1.9}$$

对于一种数值积分法，我们有必要具体地分析它的误差。因为这些误差分析有助于我们在实际计算中选到一种适宜的方法。其次，对于选定的一种方法，其 n 的选择也主要基于

这些误差分析。

关于舍入误差的分析，本章不准备讨论，读者可参阅参考资料[1]中的第四章。但我们在此指出，从理论分析和实际计算表明，舍入误差有可能随 n 的增长而增长。因此，虽然舍入误差一般是比较小的，通常是可忽略的，但在应用现代高速计算机的情况下，和号中的 n 有条件取得很大，这时舍入误差就有可能超出可忽略的范围。

下面我们将介绍几种常用的数值积分法。除特殊声明外，我们一般将假定积分区间 (a, b) 是有限区间，被积函数 $f(x)$ 是 (a, b) 上的连续函数，且它在 (a, b) 上具备所需要的可求导数的性质。

§2.2 梯形求积公式

在引言中已提到过这个方法，在此我们具体地推导它的求积公式，并分析它的截断误差。

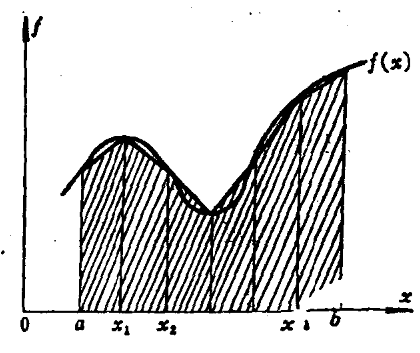


图 2.3

如图 2.3 所示，对于定积分(2.1.1)，我们用一串等分点(也可以不是等分点，但其方法原则上一样)

$$a=x_0<x_1<x_2<\cdots<x_{n-1}<x_n=b$$

将区间 $[a, b]$ 分割成 n 个小区间：

$$[x_0, x_1],\ [x_1, x_2],\ \cdots,\ [x_{n-1}, x_n]$$

每个小区间 $[x_{i-1}, x_i]$ 的长度 $h=\dfrac{b-a}{n}$，分点的坐标为

$$x_i=a+ih \quad (i=0, 1, \cdots, n)$$

在每个小区间 $[x_{i-1}, x_i]$ 上，我们用通过点 $(x_{i-1}, f(x_{i-1}))$，$(x_i, f(x_i))$ 的直线来近似函数 $f(x)$ 的弧线，即以

$$P_1(x)\equiv f(x_{i-1})+\frac{f(x_i)-f(x_{i-1})}{x_i-x_{i-1}}(x-x_{i-1})\approx f(x)\ (x_{i-1}\leqslant x\leqslant x_i) \tag{2.2.1}$$

且以此梯形面积来近似小区间上的积分。

不难求出这个梯形面积为

$$T_i=\int_{x_{i-1}}^{x_i} P_1(x)dx=\frac{f(x_{i-1})+f(x_i)}{2}h$$

以此作为小区间上积分的近似值，

$$\int_{x_{i-1}}^{x_i} f(x)dx\approx\frac{h}{2}[f(x_{i-1})+f(x_i)] \tag{2.2.2}$$

这就是梯形近似公式。

因为 $\int_a^b f(x)dx=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)dx$，因此，将这些小区间上积分的梯形近似相加起来，便得到熟知的梯形求积公式：

$$\int_a^b f(x)dx\approx h\left[\frac{f(a)}{2}+f(a+h)+f(a+2h)+\cdots+f(a+(n-1)h)+\frac{f(b)}{2}\right] \tag{2.2.3}$$

下面我们来分析梯形积分公式的截断误差 E。

如果 $f(x)$ 是一线性函数或者是顶点在 $(x_i, f(x_i))$ 处的分段线性函数，则 $f(x)$ 是与结点在 x_{i-1}，x_i 处的线性插值多项式相一致，此即 $f(x)-P_1(x)\equiv 0$，于是其截断误差

$$E=\int_{x_{i-1}}^{x_i}[f(x)-P_1(x)]dx=0$$

故梯形积分公式对线性函数(即一次多项式)是准确的

如果$f(x)$是一非线性函数，那末我们可以预见梯形积分公式的截断误差是不为零的。为此先分析在小区间上梯形近似的截断误差

$$E_i=\int_{x_{i-1}}^{x_i}[f(x)-P_1(x)]dx \tag{2.2.4}$$

假定$f(x)$在(a, b)上存在连续的二阶导数,那末根据插值多项式的余项公式可知

$$f(x)-P_1(x)=\frac{1}{2!}(x-x_{i-1})(x-x_i)f''(\bar{\xi}_i) \tag{2.2.5}$$

此处$\bar{\xi}_i$是位于x_{i-1}与x_i之间的某个值，这个值是依赖于x的，且随着x的连续变化而连续变化。因此，$\bar{\xi}_i$是x的连续函数。

将式(2.2.5)代入式(2.2.4)得

$$E_i=\frac{1}{2}\int_{x_i}^{x_{i+1}}(x-x_{i-1})(x-x_i)f''(\bar{\xi}_i)dx \tag{2.2.6}$$

根据假设条件$f''(x)$在积分区间上连续,而$\bar{\xi}_i$又是在(x_{i-1}, x_i)上的x的连续函数,因此$f''(\bar{\xi}_i)$是(x_{i-1}, x_i)上的x的连续函数。再者，$(x-x_{i-1})(x-x_i)$在所论区间上不变号，因此根据积分中值定理,在区间(x_{i-1}, x_i)内存在一个ξ_i,而有关系式

$$\frac{1}{2}\int_{x_{i-1}}^{x_i}(x-x_{i-1})(x-x_i)f''(\bar{\xi}_i)dx=\frac{1}{2}f''(\xi_i)\int_{x_{i-1}}^{x_i}(x-x_{i-1})(x-x_i)dx \tag{2.2.7}$$

因为

$$\int_{x_{i-1}}^{x_i}(x-x_{i-1})(x-x_i)dx=\frac{1}{6}(x_{i-1}-x_i)^3=-\frac{h^3}{6}$$

因此得到在第i个小区间上梯形近似的截断误差

$$E_i=-\frac{h^3}{12}f''(\xi_i)=-\frac{(b-a)^3}{12n^3}f''(\xi_i) \qquad x_{i-1}<\xi_i<x_i \tag{2.2.8}$$

将这些误差相加起来,便得到梯形求积公式的截断误差

$$E=-\frac{h^3}{12}\sum_{i=1}^{n}f''(\xi_i) \tag{2.2.9}$$

因为$f''(x)$在(a, b)上是连续的,因此在(a, b)内存在一ξ,有

$$f''(\xi)=\frac{1}{n}\sum_{i=1}^{n}f''(\xi_i) \tag{2.2.10}$$

代入式(2.2.9),于是得到

$$E=-\frac{h^2}{12}(b-a)f''(\xi)=-\frac{1}{12n^2}(b-a)^3f''(\xi) \qquad a<\xi<b \tag{2.2.11}$$

由此看出梯形求积公式的截断误差将按照$h^2\left(\text{或说按}\frac{1}{n^2}\right)$的下降速度下降。不难证明，只要$f(x)$在$(a, b)$上有界且黎曼可积，那末当分点无限增多$n\to\infty$时，梯形求积公式将收敛到积分$\int_a^b f(x)dx$。事实上，由公式(2.2.3)的等价式

$$\frac{1}{2}\sum_{i=0}^{n-1}\left(\frac{b-a}{n}\right)f(x_i)+\frac{1}{2}\sum_{i=1}^{n}\left(\frac{b-a}{n}\right)f(x_i)$$

此处$x_i=a+ih$，可以看出两和式均是形如(2.1.2)式的和式。因此，当$f(x)$黎曼可积时，它

们将趋于积分$\int_a^b f(x)dx$。这就证明了梯形求积公式的收敛性。

如上所述，梯形求积公式的基本思想是：在每个小区间上采用线性函数近似被积函数$f(x)$。如果我们不用线性函数而改用二次多项式来近似$f(x)$，那末我们就可得到在实际计算中常用的辛浦生求积公式。

§2.3 辛浦生求积公式

对定积分(2.1.1)，将积分区间$[a, b]$分割成$2n$等分，得到$2n$个小区间：

$$[x_0, x_1],\ [x_1, x_2],\ \cdots,\ [x_{2n-1}, x_{2n}],$$

此处 $x_i=a+ih \quad (i=0, 1, \cdots, 2n)$

$$h=\frac{b-a}{2n}$$

(如图2.4所示)。

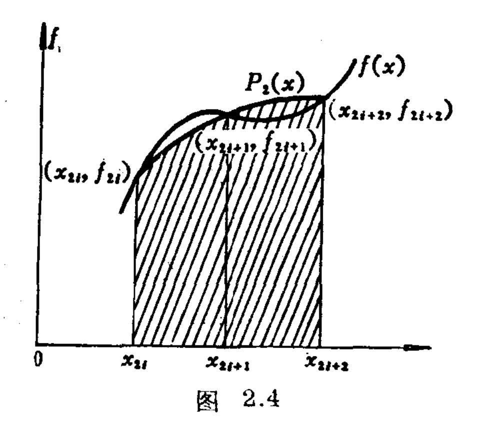


图 2.4

在每对小区间$[x_{2i}, x_{2i+1}]$，$[x_{2i+1}, x_{2i+2}]$上，用通过三个点

$$(x_{2i}, f_{2i}),\ (x_{2i+1}, f_{2i+1}),\ (x_{2i+2}, f_{2i+2})$$

(此处$f_{2i}=f(x_{2i})$，其他类同)的二次抛物线来近似函数$f(x)$的图形。即以通过上述三点的二次插值多项式$P_2(x)$来近似$f(x)$，并以$P_2(x)$在区间(x_{2i}, x_{2i+2})上的积分来近似$f(x)$在该区间上的积分

$$\int_{x_{2i}}^{x_{2i+2}} P_2(x)dx \approx \int_{x_{2i}}^{x_{2i+2}} f(x)dx \tag{2.3.1}$$

利用拉格朗日公式，我们可以写出$P_2(x)$的表达式

$$P_2(x)=\frac{1}{2h^2}[(x-x_{2i+1})(x-x_{2i+2})f_{2i}-2(x-x_{2i})(x-x_{2i+2})f_{2i+1}$$
$$+(x-x_{2i})(x-x_{2i+1})f_{2i+2}] \tag{2.3.2}$$

我们不难算出它在区间(x_{2i}, x_{2i+2})上的积分

$$\int_{x_{2i}}^{x_{2i+2}} P_2(x)dx=\frac{1}{2h^2}\left[\frac{2h^3}{3}f_{2i}+\frac{8h^3}{3}f_{2i+1}+\frac{2h^3}{3}f_{2i+2}\right] \tag{2.3.3}$$

由此便得到辛浦生近似公式

$$\int_{x_{2i}}^{x_{2i+2}} f(x)dx \approx \frac{h}{3}[f_{2i}+4f_{2i+1}+f_{2i+2}] \tag{2.3.4}$$

因为$\int_a^b f(x)dx=\sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} f(x)dx$，因此将这些小区间上积分的辛浦生近似相加起来，便得到辛浦生求积公式：

$$\int_a^b f(x)dx \approx \frac{h}{3}[f_0+4(f_1+f_3+\cdots+f_{2n-1})+2(f_2+f_4+\cdots+f_{2n-2})+f_{2n}] \tag{2.3.5}$$

此处 $f_i=f(x_i) \quad (i=0, 1, \cdots, 2n)$； $h=\frac{b-a}{2n}$； $x_i=a+ih$

利用与梯形求积公式相类似的分析方法，可以得到辛浦生求积公式的截断误差

$$E=-\frac{h^4}{180}(b-a)f^{(4)}(\xi)=-\frac{(b-a)^5}{2880n^4}f^{(4)}(\xi) \quad a<\xi<b \tag{2.3.6}$$

并可以证明辛浦生公式对于最高次数为3的多项式是准确的。

由式(2.3.6)看出，辛浦生求积公式的截断误差是按 h^4 $\left(\text{或说按}\ \frac{1}{n^4}\right)$ 的速度下降的。并易于证明：只要 $f(x)$ 在 (a, b) 上黎曼可积，那末，当分点无限增多 $n\to\infty$ 时，辛浦生求积公式将收敛到积分 $\int_a^b f(x)dx$。

梯形求积公式和辛浦生求积公式都是利用一插值多项式来近似被积函数 $f(x)$，这样的公式称为插值型积分公式。我们也可以使用更高次的插值多项式。一般地，我们可以使用一 n 次 $(n=1, 2, 3, \cdots)$ 多项式来近似被积函数 $f(x)$，这就得到熟知的牛顿-柯特斯公式。由于在实际中不常使用，此处就不加详述了。

§2.4 自动积分，逐次分半加速法

从上面介绍的几种数值积分法中，我们看到，它们的截断误差是随 n 的增长而减少的。但是对于一具体的积分问题和选定的一种数值积分法，如何确定一恰当的数 n，使得到的积分近似值 I_n 与真值 I 之间的差落在允许的误差范围之内。

自然我们可以在计算之前，根据方法的截断误差分析确定，但由于要分析被积函数的高阶导数，因此这样的做法一般是困难的。

自动积分法，是在积分计算的过程中，根据规定的精度要求，自动地确定数 n，并算出满足精度要求的积分近似值，而不需事先进行人工分析的工作。

2.4.1 基于梯形和辛浦生公式的自动积分法

由梯形求积公式的截断误差式(2.2.11)，我们看到，当取 n 时的截断误差是

$$E_n=I-I_n=-\frac{(b-a)^3}{12n^2}f''(\xi_n) \quad a<\xi_n<b$$

当取 $2n$ 时的截断误差是

$$E_{2n}=I-I_{2n}=-\frac{(b-a)^3}{12(2n)^2}f''(\xi_{2n}) \quad a<\xi_{2n}<b$$

此处 I 是积分真值。

将上两式相减得

$$I_{2n}-I_n=-\frac{(b-a)^3}{12(2n)^2}[4f''(\xi_n)-f''(\xi_{2n})]$$

且从 $f''(\xi)$ 的含义(2.2.10)式可知，当 n 充分大时有 $f''(\xi_n)\approx f''(\xi_{2n})$，故得

$$\frac{1}{3}(I_{2n}-I_n)\approx I-I_{2n} \tag{2.4.1}$$

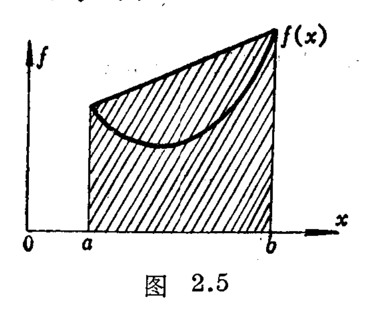


图 2.5

式(2.4.1)提供了一个方便的误差判据。我们可以由检验条件

$$|I_{2n}-I_n|<\varepsilon\text{（允许误差）}$$

来判断积分近似值 I_{2n} 是否已满足精度要求。由此我们就可以构造下述自动积分过程。

我们采用将积分区间逐次分半的分割法，计算梯形和序列，即

最初取 $n=1$（如图2.5），计算

$$I_1=h_0\left[\frac{f(a)}{2}+\frac{f(b)}{2}\right]\quad h_0=b-a$$

然后将区间分割为 2 (如图 2.6), 取 $n=2$, 计算

$$I_2=h_1\left[\frac{f(a)}{2}+\frac{f(b)}{2}+f(x_1)\right]=\frac{I_1}{2}+h_1f(x_1)$$

$$h_1=\frac{b-a}{2}\qquad x_1=a+h_1$$

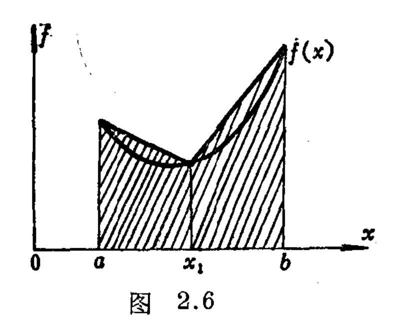


图 2.6

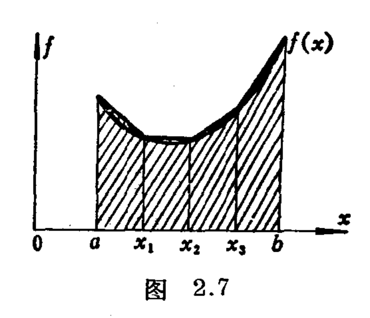


图 2.7

然后将区间分割为 4 (如图 2.7), 取 $n=4$, 计算

$$I_4=h_2\left[\frac{f(a)}{2}+\frac{f(b)}{2}+f(x_1)+f(x_2)+f(x_3)\right]=\frac{I_2}{2}+h_2[f(x_1)+f(x_3)]$$

$$h_2=\frac{b-a}{4},\quad x_i=a+ih_2(i=1,\ 2,\ 3)$$

一般地, 每次总是在前一次的基础上再将小区间分半, 分点加密一倍。而老分点上的函数值便不需再重复计算。其一般算式是:

$$\left.\begin{aligned}I_{2n}&=\frac{I_n}{2}+h_{2n}\sum_{i=1}^{n}f(a+(2i-1)h_{2n})\\ h_{2n}&=\frac{b-a}{2n}\end{aligned}\right\}\tag{2.4.2}$$

在计算梯形和序列的过程中, 每当算出一新的近似值 I_{2n} 时, 使用检验条件:

$$\left.\begin{aligned}&|I_{2n}-I_n|<\varepsilon\ (\text{取绝对误差时})\\ \text{或}\qquad &\frac{|I_{2n}-I_n|}{|I_{2n}|}<\varepsilon\ (\text{取相对误差时})\end{aligned}\right\}\tag{2.4.3}$$

当条件满足时, 由式(2.4.1)知, I_{2n} 即是符合精度要求的积分近似值。

在实际使用中, 为预防假收敛, 一般还需给出 n 的下界数 $\min n$。当满足条件(2.4.3)且 $2n>\min n$ 时, 方认为计算收敛。因为对于振荡的被积函数, 有可能发生假收敛。例如当被积函数具有如图 2.8 所示的图形时, 因 $I_2-I_1=0$ 而就会产生假收敛。

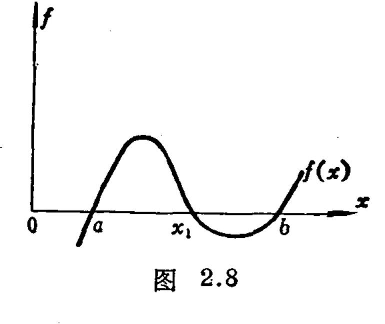


图 2.8

关于方法的收敛速度, 不难得到

$$I_{2n}-I_n\approx\frac{1}{4}(I_n-I_{\frac{n}{2}})\tag{2.4.4}$$

因此是收敛因子为 1/4 的线性收敛速度。

基于辛浦生公式的自动积分法也是类似的。

从辛浦生公式的截断误差, 我们不难得到关系式

$$\frac{1}{15}(I_{2n}-I_n)\approx(I-I_{2n})\tag{2.4.5}$$

因此可以由 $I_{2n}-I_n$ 来估计近似值 I_{2n} 的误差。

同样采用区间逐次分半的分割法，计算辛浦生积分序列：$I_1, I_2, I_4, \cdots, I_n, I_{2n}, \cdots$，

$$I_n=\frac{h_n}{3}[f(a)+f(b)+2S_n+4\bar{S}_n], n=1, 2, 4, \cdots$$

此处，S_n 是在老分点上函数值之和，

$$S_n=f(x_2)+f(x_4)+\cdots+f(x_{2n-2})$$

$\bar{S}_n$ 是在新分点上函数值之和，

$$\bar{S}_n=f(x_1)+f(x_3)+\cdots+f(x_{2n-1})$$

$$h_n=\frac{b-a}{2n}, \ x_i=a+ih_n \ (i=1, \cdots, 2n-1)$$

在老分点上的函数值每次不需重复计算。

同样地，每次检验条件(2.4.3)，若条件满足且 $2n>\min n$ 时，I_{2n} 便是满足精度要求的积分近似值。

关于方法的收敛速度，我们不难得到

$$I_{2n}-I_n\approx\frac{1}{16}(I_n-I_{\frac{n}{2}}) \tag{2.4.6}$$

因此是以 $\frac{1}{16}$ 为收敛因子的线性收敛速度。

2.4.2 逐次分半加速法(Romberg 积分法)

逐次分半加速法，亦称龙勃(Romberg)积分法，它是在计算梯形和序列的基础上，应用了线性外插的加速方法，由此构成一种具有超线性收敛速度的自动积分法。该方法的思想如下：

按照区间逐次分半的方法，计算梯形和序列

$$n=1, \quad h_0=b-a, \quad T_{00}=h_0\left[\frac{f(a)}{2}+\frac{f(b)}{2}\right]$$

$$n=2, \quad h_1=\frac{b-a}{2}, \quad T_{01}=h_1\left[\frac{f(a)}{2}+\frac{f(b)}{2}+f(a+h_1)\right]$$

$$n=2^2, \quad h_2=\frac{b-a}{2^2}, \quad T_{02}=h_2\left[\frac{f(a)}{2}+\frac{f(b)}{2}+\sum_{i=1}^{3}f(a+ih_2)\right]$$

$$\cdots\cdots$$

$$n=2^k, \quad h_k=\frac{b-a}{2^k}, \quad T_{0k}=h_k\left[\frac{f(a)}{2}+\frac{f(b)}{2}+\sum_{i=1}^{2^k-1}f(a+ih_k)\right] \tag{2.4.7}$$

$$\cdots\cdots$$

在 §2.2 中已经证明了，只要 $f(x)$ 在 $[a, b]$ 上有界且黎曼可积，那末梯形和序列收敛到积分 $I=\int_a^b f(x)dx$。并且，如果 $f(x)$ 在 (a, b) 上存在二阶连续导数，那末根据梯形求积公式的截断误差有

$$T_0=I+\frac{(b-a)}{12}f''(\xi)h^2, \quad a<\xi<b \tag{2.4.8}$$

此处 I 是积分真值；ξ 是和 h 有关的量，可视为 h 的函数。

式(2.4.8)也反映了梯形和 T_0 与步长 h^2 之间的关系。如图 2.9 所示，如果我们将 h^2 作

为自变量，将 T_0 视为 h^2 的函数，那末在以 h^2 为横轴，以 T_0 为纵轴的平面上，可以画出一条曲线，这条曲线在 T_0 轴的截距即是积分真值 I。而梯形和序列的值对：

$(h_0^2,\ T_{00}),\ (h_1^2,\ T_{01}),\ \cdots,\ (h_k^2,\ T_{0k}),\ \cdots$

是位于曲线图形上的一串点列，这串点列以点 $(0,\ I)$ 为极限。

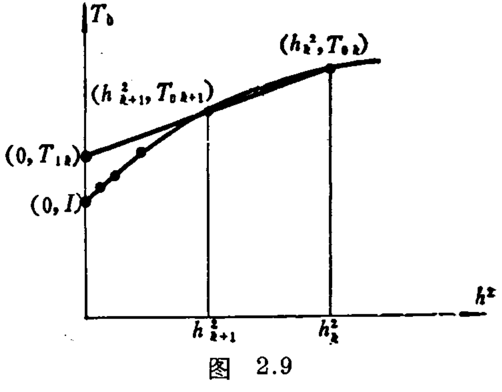


图 2.9

对于梯形和序列中的任何二相邻元素 T_{0k}，T_{0k+1}，相应于图形上的两个点。如果我们通过这两个点作一直线，并将此直线外延到 T_0 轴，得到一截距 T_{1k}，那末这个 T_{1k} 可以期望是比 T_{0k+1} 更为接近于 I。这就是从几何上看的线性外插加速的思想。

龙勃方法就是从这个基本思想出发所构成的一种方法。下面就来导出这个算法。

设 T_{0k} 和 T_{0k+1} 分别是以 h_k 和 h_{k+1} 为步长的梯形和。我们不难写出通过点 $(h_k^2,\ T_{0k})$ 和 $(h_{k+1}^2,\ T_{0k+1})$ 的直线方程式

$$T_0=T_{0k}+\frac{T_{0k+1}-T_{0k}}{h_{k+1}^2-h_k^2}(h^2-h_k^2)$$

其在 T_0 轴上的截距是

$$T_{1k}=T_{0k}-\frac{T_{0k+1}-T_{0k}}{h_{k+1}^2-h_k^2}h_k^2$$

利用 $h_{k+1}=\dfrac{h_k}{2}$ 的关系，便得到新近似值

$$T_{1k}=\frac{4T_{0k+1}-T_{0k}}{4-1} \tag{2.4.9}$$

我们可以证明公式(2.4.9)是等价于以 h_{k+1} 为步长的辛浦生求积公式。而辛浦生求积公式是具有 h^4 量级的截断误差，即有

$$T_{1k}=I+\frac{(b-a)}{180}f^{(4)}(\xi_k)h_{k+1}^4,\quad a<\xi_k<b \tag{2.4.10}$$

因此这个新近似值 T_{1k} 确实是比 T_{0k+1} 更接近于 I。

如果对梯形和序列的每相邻两元素均应用公式 (2.4.9) 就可以得到 T_{1k} $(k=0,\ 1,\ 2,\ \cdots)$ 序列。

由式(2.4.10)，类似地，如果我们以 h^4 为自变量，视 T_{1k} 为 h^4 的函数，再应用线性外插过程，可以得到又一新的序列，T_{2k} 序列

$$T_{2k}=\frac{4^2T_{1k+1}-T_{1k}}{4^2-1},\quad k=0,\ 1,\ 2,\ \cdots \tag{2.4.11}$$

可以证明公式(2.4.11)是等价于以 h_{k+2} 为步长的四阶牛顿-柯特斯公式(即用四次插值多项式近似被积函数后导出的公式)，具有 h_{k+2}^6 量级的截断误差

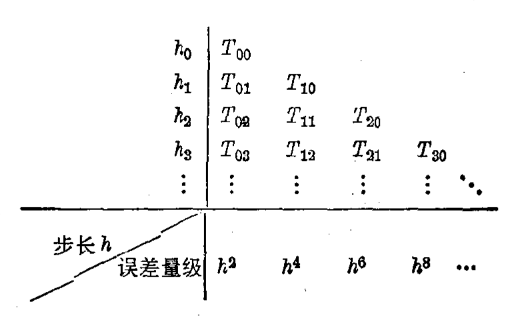
$$T_{2k}=I+\frac{2(b-a)}{945}f^{(6)}(\xi_k)h_{k+2}^6 \tag{2.4.12}$$

同样，对序列 $T_{2k}(k=0,\ 1,\ 2,\ \cdots)$ 又可执行对变量 h^6 的线性外插过程得 $T_{3k}(k=0,\ 1,\ 2,\ \cdots)$ 序列。如此继续进行。每当执行一次线性外插，误差的量级便增加一个 h^2。一般地，由 $T_{m-1k}(k=0,\ 1,\ 2,\ \cdots)$ 序列可执行对变量 h^{2m} 的线性外插过程，得序列

$$T_{mk}=\frac{4^m T_{m-1\,k+1}-T_{m-1\,k}}{4^m-1} \tag{2.4.13}$$

这个公式具有 $h_{k+m}^{2(m+1)}$ 量级的截断误差。

按此作法过程，可以得到如下的三角形数表，称之为 T 表：



h_0	T_{00}				
h_1	T_{01}	T_{10}			
h_2	T_{02}	T_{11}	T_{20}		
h_3	T_{03}	T_{12}	T_{21}	T_{30}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
步长 h / 误差量级	h^2	h^4	h^6	h^8	…

对于这样的 T 表，在资料[3]中已经证明了：

(1) 如果 $f(x)$ 在 $[a, b]$ 上有界且是黎曼可积的，那末 T 表的列和对角线都收敛到积分 I。

(2) 如果 $f(x)$ 在 $[a, b]$ 上是解析的，则 T 表的任何对角线便按超线性的速度收敛到 I。

为使用上的明确起见，下面我们把龙勃方法的具体计算步骤简列于下：

(1) 取 $h_0=b-a$，计算 $S_0=\frac{1}{2}[f(a)+f(b)]$ 和梯形和 $T_{00}=h_0S_0$；

对 $k=1, 2, \cdots$，计算下列各步：

(2) 取 $h_k=h_{k-1}/2$，计算 $S_k=S_{k-1}+\sum\limits_{i=1}^{2^{k-1}}f(a+(2i-1)h_k)$ 及梯形和 $T_{0k}=h_kS_k$；

(3) 作线性外插计算，算出 T 表的第 k 行元素，即对 $m=1, 2, \cdots, k$，相应地 $j=k-1, k-2, \cdots$，

计算 k 个量
$$T_{mj}=\frac{T_{m-1\,j+1}-4^{-m}T_{m-1\,j}}{1-4^{-m}}$$

将它们连同 T_{0k} 一起保存在一维场 $T=\{T_0, T_1, \cdots, T_{\max k}\}$ 中，其次序为：

$$\begin{array}{lccccccc} T: & T_0 & T_1 & T_2 & \cdots & T_k & \cdots & T_{\max k} \\ & \uparrow & \uparrow & \uparrow & & \uparrow & & \\ \text{保存的量:} & T_{k0} & T_{k-1\,0} & T_{k-2\,0} & \cdots & T_{0k} & & \end{array}$$

(4) 收敛控制

如果 T 表对角线上的最后两相邻元素 T_{k0}、$T_{k-1\,0}$ 满足

$$|T_{k0}-T_{k-1\,0}|<\varepsilon\text{(取绝对误差时)}$$

或
$$\frac{|T_{k0}-T_{k-1\,0}|}{|T_{k0}|}<\varepsilon\text{(取相对误差时)}$$

并且 $k\geqslant\min k$，则以 T_{k0} 作为积分近似值，完成求积计算，否则继续到步骤(5)。此处 $\min k$ 是最低限度的区间分半次数，起防止假收敛的作用。

(5) 若 $k<\max k$，则 k 增加 1 后转到步骤(2)继续计算，否则计算不成功，终止计算。此处 $\max k$ 是最高限度的区间分半次数，起控制工作量的作用。

在§2.2，§2.3中，我们提到了梯形公式对于最高次数为 1 次的多项式是准确的；辛浦生公式对于最高次数为 3 次的多项式是准确的。如果一个积分公式对于次数不超过 m 次的多项式是准确的，而对 $m+1$ 次多项式不再准确的话，那末便称此积分公式具有 m 次的

代数准确度。

一个积分公式的代数准确度次数也是该积分公式近似程度的一种量度。对于包含同样结点数的积分公式，从经济的角度，自然我们希望采用一种准确度次数比较高的积分公式，因为这样可以用同等的代价，获得较高近似度的计算结果。

因此，在建立积分公式的工作中，不免提出这样的问题：我们能不能适当地选择 n 个结点和相应的 n 个系数，使得积分公式具有最大的代数准确度。从这个观点导出的积分公式便是下面要介绍的高斯型积分公式。

§2.5 高斯型求积公式

为了一般性，考虑积分

$$I=\int_a^b W(x)f(x)dx \tag{2.5.1}$$

其中 $W(x)$ $(W(x)\geqslant 0)$ 称为权函数。当取 $W(x)=1$ 时，即是普通的积分。对于任何普通积分 $\int_a^b f(x)dx$，都可写成 $\int_a^b W(x)\dfrac{f(x)}{W(x)}dx$，从而都可化成此种形式的积分。

对于积分(2.5.1)，假定我们采用具有 n 个结点的积分公式：

$$\int_a^b W(x)f(x)dx\approx\sum_{i=1}^n c_i f(x_i) \tag{2.5.2}$$

其中系数 $c_i(i=1, 2, \cdots, n)$ 不依赖于函数 $f(x)$，但可以依赖于权函数 $W(x)$。目的是适当地选择 n 个结点 $x_1, x_2, \cdots, x_n$ 和相应的 n 个系数 $c_1, c_2, \cdots, c_n$，使得积分公式(2.5.2)具有最大次数的代数准确度。

首先考察对于固定的 n 值，公式(2.5.2)最大可以达到多少次的代数准确度。

假定公式(2.5.2)对所有的 m (m 待定)次多项式

$$P_m(x)=a_mx^m+a_{m-1}x^{m-1}+\cdots+a_1x+a_0$$

是准确的。于是有

$$\begin{aligned}a_m\int_a^b W(x)x^m dx+a_{m-1}\int_a^b W(x)x^{m-1}dx+\cdots+a_1\int_a^b W(x)x\,dx+a_0\int_a^b W(x)dx\\=\sum_{i=1}^n c_i(a_mx_i^m+a_{m-1}x_i^{m-1}+\cdots+a_1x_i+a_0)\end{aligned} \tag{2.5.3}$$

令 $\mu_k=\int_a^b W(x)x^k dx(k=0, 1, \cdots, m)$，并重新组合(2.5.3)式右端各项，得

$$\begin{aligned}&a_m\mu_m+a_{m-1}\mu_{m-1}+\cdots+a_1\mu_1+a_0\mu_0\\&\quad=a_m\sum_{i=1}^n c_ix_i^m+a_{m-1}\sum_{i=1}^n c_ix_i^{m-1}+\cdots+a_1\sum_{i=1}^n c_ix_i+a_0\sum_{i=1}^n c_i\end{aligned} \tag{2.5.4}$$

由于系数 $a_m, \cdots, a_0$ 的任意性，使式(2.5.4)成立的充要条件是

$$\left.\begin{aligned}c_1+c_2+\cdots+c_n&=\mu_0\\c_1x_1+c_2x_2+\cdots+c_nx_n&=\mu_1\\c_1x_1^2+c_2x_2^2+\cdots+c_nx_n^2&=\mu_2\\\cdots\cdots&\\c_1x_1^m+c_2x_2^m+\cdots+c_nx_n^m&=\mu_m\end{aligned}\right\} \tag{2.5.5}$$

因为 $2n$ 个待定数最多只能给 $2n$ 个独立的条件，因此可知 m 最多为 $2n-1$。

由此得出，对于 n 个结点的积分公式，其最大可能的代数准确度次数是 $2n-1$。并且可以证明，方程(2.5.5)当取 $m=2n-1$时是可解的。因此，确实可以找到一组 x_i 和 $c_i(i=1, 2, \cdots, n)$，使积分公式(2.5.2)达到 $2n-1$ 次的代数准确度。这样的公式就是高斯型求积公式。

关于高斯型求积公式的结点和系数，可以从方程(2.5.5)解得，但一般是利用正交多项式来确定它们。

我们知道：

如果两个多项式 $Q_i(x)$，$Q_j(x)$（下标表示多项式的次数）有

$$\int_a^b W(x)Q_i(x)Q_j(x)dx=0 \tag{2.5.6}$$

那末称多项式 $Q_i(x)$ 和 $Q_j(x)$ 在区间 $[a, b]$ 上关于权 $W(x)$ 为正交。

如果一多项式序列 $Q_0(x)$, $Q_1(x)$, $\cdots$, $Q_n(x)$, $\cdots$具有性质：

$$\int_a^b W(x)Q_i(x)Q_j(x)dx=0 \quad (i,\ j=1,\ 2,\ \cdots,\ i\neq j) \tag{2.5.7}$$

则称此多项式序列 $\{Q_n(x)\}$ 为 $[a, b]$ 上关于权函数 $W(x)$ 的正交多项式系。如果进一步有：

$$\int_a^b W(x)Q_i^2(x)=1 \quad (i=1,\ 2,\ \cdots)$$

则称 $\{Q_n(x)\}$ 为 $[a, b]$ 上关于权函数 $W(x)$ 的规格化正交多项式系。

对于 $[a, b]$ 上关于非负权函数 $W(x)$ 的正交多项式系 $\{Q_n(x)\}$，其 $Q_n(x)$ 的 n 个零点是实的，不相重的，且分布在开区间 (a, b) 之中。并且，对于一给定的权函数，总存在关于此权函数的正交多项式系。

利用正交多项式的一些关系式和性质，可以导出：

高斯型求积公式(2.5.2)的 n 个结点 $x_1, x_2, \cdots, x_n$ 是 $[a, b]$ 上关于权 $W(x)$ 的 n 次正交多项式 $Q_n(x)$ 的 n 个零点；

高斯型求积公式的 n 个系数为：

$$c_i=\frac{1}{Q_n^{*\prime}(x_i)}\int_a^b \frac{W(x)Q_n^*(x)}{(x-x_i)}dx=\frac{a_n}{a_{n-1}Q_n^{*\prime}(x_i)Q_{n-1}^*(x_i)} \quad (i=1,\ 2,\ \cdots,\ n) \tag{2.5.8}$$

此处 $Q_n^*(x)$ 是 $[a, b]$ 上关于权 $W(x)$ 的 n 次规格化正交多项式，a_n 和 a_{n-1} 分别是 n 次和 $n-1$ 次规格化正交多项式 $Q_n^*(x)$ 和 $Q_{n-1}^*(x)$ 的首项系数。

下面讨论高斯型积分公式的截断误差：

$$E=\int_a^b W(x)f(x)dx-\sum_{i=1}^n c_i f(x_i) \tag{2.5.9}$$

假设 $f(x)$ 在 (a, b) 上存在 $2n$ 阶连续导数。利用在结点 $x_1, x_2, \cdots, x_n$ 上的埃尔米特插值公式，可写

$$f(x)=H_{2n-1}(x)+\frac{f^{(2n)}(\xi)}{(2n)!}(x-x_1)^2\cdots(x-x_n)^2 \quad a<\xi<b \tag{2.5.10}$$

此处 $H_{2n-1}(x)$ 是 $2n-1$ 次的埃尔米特插值多项式。将(2.5.10)代入(2.5.9)得

$$E=\int_a^b W(x)H_{2n-1}(x)dx+\frac{1}{(2n)!}\int_a^b W(x)f^{(2n)}(\xi)\omega_n^2(x)dx-\sum_{i=1}^n c_i H_{2n-1}(x_i)$$

式中 $\omega_n(x)=(x-x_1)\cdots(x-x_n)$。因高斯型求积公式对任何 $2n-1$ 次的多项式是准确的，故得

$$E=\frac{1}{(2n)!}\int_a^b W(x)f^{(2n)}(\xi)\omega_n^2(x)dx=\frac{1}{a_n^2(2n)!}\int_a^b W(x)f^{(2n)}(\xi)Q_n^2(x)dx$$

式中 $Q_n(x)=a_n\omega_n(x)$ 是 $[a,\ b]$ 上关于权 $W(x)$ 的正交多项式；a_n 是其首项系数。

再利用积分中值定理，便得

$$E=\frac{f^{(2n)}(\xi)}{a_n^2(2n)!}\int_a^b W(x)Q_n^2(x)dx \quad a<\xi<b \tag{2.5.11}$$

此外，可以证明只要 $f(x)$ 在 $(a,\ b)$ 上连续，那末当 $n\to\infty$ 时，高斯型积分公式收敛于定积分，即有

$$\lim_{n\to\infty}\sum_{i=1}^{n}c_i f(x_i)=\int_a^b W(x)f(x)dx$$

上面，对于一般的权函数讨论了高斯型积分公式。对于不同的权函数，便得到不同的具体的高斯型积分公式。下面，对于两种具体的权函数给出相应的高斯积分公式。

不失一般性，假定积分区间 $(a,\ b)$ 是 $(-1,\ 1)$，因为总可以利用积分变量的变换

$$x=\frac{b+a}{2}+\frac{b-a}{2}t$$

将区间 $(a,\ b)$ 变成 $(-1,\ 1)$，而积分变为

$$\int_a^b f(x)dx=\frac{b-a}{2}\int_{-1}^{1}g(t)dt$$

式中

$$g(t)=f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)$$

(一) 高斯-勒让德求积公式

高斯-勒让德求积公式是相应于权函数 $W(x)=1$ 时的积分公式，它是古典的高斯求积公式，故一般就称之为高斯求积公式。

我们知道勒让德多项式

$$L_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n]=\sum_{k=0}^{\left[\frac{n}{2}\right]}(-1)^k\frac{(2n-2k)!}{2^n\cdot k!\cdot(n-k)!\cdot(n-2k)!}x^{n-2k} \tag{2.5.12}$$

$$(n=1,\ 2,\ \cdots)\quad \left(\left[\frac{n}{2}\right]\text{表示}\ \frac{n}{2}\ \text{的整数部分}\right)$$

是构成在区间 $[-1,\ 1]$ 上的正交系，即可以证明

$$\int_{-1}^{1}L_m(x)L_n(x)dx=\begin{cases}0, & (m\neq n)\\ \dfrac{2}{2n+1}, & (m=n)\end{cases}\quad (m,\ n=1,\ 2,\ \cdots) \tag{2.5.13}$$

且不难得到规格化的勒让德多项式

$$L_n^*(x)=\sqrt{\frac{2n+1}{2}}L_n(x)=\frac{(2n)!}{2^n(n!)^2}\sqrt{\frac{2n+1}{2}}x^n+\cdots \tag{2.5.14}$$

因此根据前面介绍的高斯型积分公式结点和系数的确定法，可知高斯-勒让德求积公式中的 n 个结点 $x_1,\ x_2,\ \cdots,\ x_n$ 就是 n 次勒让德多项式 $L_n(x)$ 的 n 个零点，而系数 c_i 按公式 (2.5.8) 再利用勒让德多项式的一个性质

$$(1-x^2)L_n'(x)=n[L_{n-1}(x)-xL_n(x)] \tag{2.5.15}$$

可得

$$\left.\begin{aligned} c_i&=\frac{2}{(1-x_i^2)[L_n'(x_i)]^2} \\ c_i&=\frac{2(1-x_i^2)}{[nL_{n-1}(x_i)]^2}\end{aligned}\right\} i=1,\ 2,\ \cdots,\ n \tag{2.5.16}$$

或

显然这些系数 c_i 都是正数。

关于高斯-勒让德求积公式的截断误差，按(2.5.11)式，不难得到

$$E_n=\frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3}f^{(2n)}(\xi) \quad -1<\xi<1 \tag{2.5.17}$$

在实际使用上，勒让德多项式的计算可以利用三次循环公式：

$$\left.\begin{aligned} &L_0(x)=1 \\ &L_1(x)=x \\ &nL_n(x)=(2n-1)xL_{n-1}(x)-(n-1)L_{n-2}(x) \quad n\geqslant 2\end{aligned}\right\} \tag{2.5.18}$$

勒让德多项式的零点 $x_i(i=1,\ 2,\ \cdots,\ n)$ 可以用求函数零点的牛顿迭代法(见第九章)：

$$x_i^{k+1}=x_i^k-\frac{L_n(x_i^k)}{L_n'(x_i)}$$

来逐个地求出。其中 $L_n'(x_i^k)$ 可以利用关系式(2.5.15)来计算。

在本章的附表1中，具体给出了 $n=2$ 到 20 时的高斯-勒让德求积公式的结点和系数。

(二)高斯-切比雪夫求积公式

高斯-切比雪夫求积公式是相应于权函数 $W(x)=1/\sqrt{1-x^2}$ 时的高斯型积分公式。

我们知道在 $[-1,\ 1]$ 上关于权函数 $1/\sqrt{1-x^2}$ 的正交多项式是切比雪夫多项式

$$\left.\begin{aligned} T_n(x)&=\cos(n\cos^{-1}x) \\ &=2^{n-1}x^n+\cdots \quad n=1,\ 2,\ \cdots \\ T_0(x)&=1\end{aligned}\right\} \tag{2.5.19}$$

因为可以证明

$$\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}T_m(x)T_n(x)dx=\begin{cases}0, & m\neq n \\ \dfrac{\pi}{2}, & m=n\neq 0 \\ \pi, & m=n=0\end{cases} \quad (m,\ n=1,\ 2,\ \cdots) \tag{2.5.20}$$

且不难得到规格化的切比雪夫多项式

$$\left.\begin{aligned} T_n^*(x)&=\sqrt{\frac{2}{\pi}}\ T_n(x)=\sqrt{\frac{2}{\pi}}\ 2^{n-1}x^n+\cdots \\ T_0^*(x)&=\sqrt{\frac{1}{\pi}}\end{aligned}\right\} \tag{2.5.21}$$

因此根据高斯型积分公式结点和系数的确定法，可知高斯-切比雪夫积分公式中的结点 $x_1,\ x_2,\ \cdots,\ x_n$ 就是 n 次切比雪夫多项式 $T_n(x)$ 的 n 个零点

$$x_i=\cos\theta_i=\cos\left(\frac{2i-1}{2n}\pi\right) \quad i=1,\ 2,\ \cdots,\ n \tag{2.5.22}$$

而系数 c_i 按公式(2.5.8)得

$$c_i=\frac{\pi}{T_n'(x_i)T_{n-1}(x_i)}$$

根据 $T'_n(x_i)=\frac{n\sin n\theta_i}{\sin\theta_i}$ 并利用关系式: $\cos(n-1)\theta_i=\sin\theta_i\sin n\theta_i$ 和 $\sin\theta_i=1$, 可得

$$c_i=\frac{\pi}{n}\quad i=1,\ 2,\ \cdots,\ n \tag{2.5.23}$$

于是导出高斯-切比雪夫积分公式

$$\left.\begin{aligned}&\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}f(x)dx\approx\frac{\pi}{n}\sum_{i=1}^{n}f(x_i)\\ &x_i=\cos\left(\frac{2i-1}{2n}\pi\right)\quad i=1,\ 2,\ \cdots,\ n\end{aligned}\right\} \tag{2.5.24}$$

其中

高斯-切比雪夫积分公式的截断误差:

由(2.5.11)和(2.5.19), (2.5.20)可得

$$E=\frac{\pi}{2^{2n-1}(2n)!}f^{(2n)}(\xi)\quad -1<\xi<1 \tag{2.5.25}$$

§2.6 用切氏级数展开的积分法及方法比较

在实际问题中,除了定积分的计算外,也常遇到更为一般的问题,即计算不定积分

$$I(x)=\int_a^x f(x)dx\quad a\leqslant x\leqslant b \tag{2.6.1}$$

对于这样的问题,自然也可以在 x 的变化范围内取定一串数值,将不定积分的计算化为一串定积分的计算。但是,这样只能得到若干离散点上的积分值,并且取多少个 x 值就得计算多少个定积分。

下面将要介绍的用切氏级数展开的积分法, 却能比较好地解决这个问题。因为它可以给出不定积分的一个近似的函数表达式。这样,对于在$[a,\ b]$上的任意的 x 值, 可以很方便地由计算函数表达式来得到积分 $I(x)$的近似值,而不需每次进行定积分的计算。

在介绍这个方法之前, 有必要将所需的关于切比雪夫多项式和函数的切氏级数展开的有关知识简单地提一下。

(1) 切比雪夫多项式是由下式所定义

$$T_n(x)=\cos(n\cos^{-1}x)=2^{n-1}x^n+\cdots\quad n=1,\ 2\ \cdots;\ -1\leqslant x\leqslant 1 \tag{2.6.2}$$

它满足如下的递推关系式

$$T_{n+1}(x)=2xT_n(x)-T_{n-1}(x)\quad n=1,\ 2,\ \cdots \tag{2.6.3}$$

而

$$T_0(x)=1$$

$$T_1(x)=x$$

切比雪夫多项式是有界的:

$$|T_n(x)|\leqslant 1\quad -1\leqslant x\leqslant 1 \tag{2.6.4}$$

切比雪夫多项式除了由式(2.5.20)所示的正交性质外,还有在如下意义上的正交性

$$\sum_{k=0}^{N}{}''T_m(x_k)T_n(x_k)=\begin{cases}N, & m=n=0 \text{ 或 } N\\ N/2, & m=n \text{ 但 } \neq 0 \text{ 或 } N\\ 0, & m\neq n\end{cases} \tag{2.6.5}$$

$$(m,\ n\leqslant N)$$

此处

$$x_k=\cos\frac{k\pi}{N}\quad k=0,\ 1,\ \cdots,\ N$$

Σ'' 表示求和时对 $k=0$ 和 N 的项取半(即乘以 1/2)。

切比雪夫多项式的不定积分是切比雪夫多项式本身的线性组合,即有:

$$\left.\begin{aligned}&\int_{-1}^{x} T_0(x)\,dx = T_1(x)+1\\&\int_{-1}^{x} T_1(x)\,dx = \frac{1}{4}T_2(x)-\frac{1}{4}\\&\int_{-1}^{x} T_n(x)\,dx = \frac{1}{2}\left(\frac{T_{n+1}(x)}{n+1}-\frac{T_{n-1}(x)}{n-1}\right)+(-1)^{n-1}\frac{1}{n^2-1}\quad n=2,3,\cdots\end{aligned}\right\}-1\leqslant x\leqslant 1 \quad (2.6.6)$$

证明:对于 $n=0,\ 1$, 不难分别验证。对于 $n>1$, 我们有

$$\begin{aligned}\int_{-1}^{x} T_n(x)\,dx &= \int_{-1}^{x}\cos(n\cos^{-1}x)\,dx = -\int_{x}^{\cos^{-1}x}\cos n\theta\sin\theta\,d\theta\\&=\left[\frac{\cos(n+1)\theta}{n+1}-\frac{\cos(n-1)\theta}{n-1}\right]_{x}^{\cos^{-1}x}\\&=\frac{1}{2}\left(\frac{T_{n+1}(x)}{n+1}-\frac{T_{n-1}(x)}{n-1}\right)+\frac{1}{2}\left[\frac{(-1)^{n-1}}{n-1}-\frac{(-1)^{n+1}}{n+1}\right]\end{aligned}$$

(2) 对于任何$[-1,\ 1]$上的有界变差的连续函数 $f(x)$, 可以展成一致收敛的切比雪夫多项式的级数

$$f(x)=\frac{1}{2}A_0T_0(x)+A_1T_1(x)+A_2T_2(x)+\cdots \quad (2.6.7)$$

式中 $T_k(x)\ (k=0,\ 1,\ \cdots)$ 是切比雪夫多项式,而系数 A_k, 利用正交关系式(2.5.20)容易得到

$$A_k=\frac{2}{\pi}\int_{-1}^{1}\frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,dx=\frac{2}{\pi}\int_{0}^{\pi}f(\cos\theta)\cos k\theta\,d\theta \quad (k=0,\ 1,\ \cdots) \quad (2.6.8)$$

当 k 足够大时,这些系数 $A_k(k=0,\ 1,\ \cdots)$ 至少按 k^{-2} 的速度下降趋于零。并且,如果 $f(x)$ 在$[-1,\ 1]$上存在连续的 r 阶导数,则当 k 足够大时,A_{2k} 将按 $k^{-(r+1)}$ 的下降速度趋于零;如果 $f(x)$ 在一包含实轴段$[-1,\ 1]$的区域上解析,则 A_{2k} 将按 $e^{-\alpha k}$(α 是某个>0 的数)的下降速度趋于零。

如果取级数(2.6.7)的部分和作为 $f(x)$ 的近似式

$$f(x)\approx S_N(x)=\frac{1}{2}A_0+A_1T_1(x)+\cdots+A_NT_N(x) \quad (2.6.9)$$

这个部分和是一次数$\leqslant N$ 的多项式,这个多项式非常接近于在 N 次多项式 $P_N(x)$ 中使偏差 $\max\limits_{-1\leqslant x\leqslant 1}|f(x)-P_N(x)|$ 为最小的 $f(x)$ 的最优逼近多项式。

现在回到所要介绍的用切氏级数展开的积分法。

不失一般性,假定式(2.6.1)中的积分区间$[a,\ b]$是$[-1,\ 1]$。因为对于任何的区间$[a,\ b]$,我们总可以施以积分变量的线性变换,将区间变到$[-1,\ 1]$。

用切氏级数展开的积分法是将被积函数 $f(x)$ 展成切氏级数,然后取级数的部分和

$$S_N(x)=\frac{1}{2}A_0+A_1T_1(x)+\cdots+A_NT_N(x) \quad (2.6.10)$$

来近似 $f(x)$,其中 $A_k(k=0,\ 1,\ \cdots,\ N)$ 是由(2.6.8)式给出。

但要从(2.6.8)式来计算这些系数一般是困难的,通常是采用 $S_N(x)$ 的一种近似式:

$$\bar{S}_N(x)=\frac{a_0}{2}+a_1T_1(x)+\cdots+a_{N-1}T_{N-1}(x)+\frac{a_N}{2}T_N(x) \quad (2.6.11)$$

其中系数 $a_k(k=0,\ 1,\ \cdots,\ N)$ 是由下式给出

$$a_k=\frac{2}{N}\sum_{j=0}^{N}{}'' f\left(\cos\frac{j\pi}{N}\right)\cos\frac{kj\pi}{N} \tag{2.6.12}$$

此处符号 Σ'' 表示对 $j=0$ 和 N 的项取半(即乘以 1/2)。这个近似式 $\bar{S}_N(x)$，事实上是以

$$x_j=\cos\frac{j\pi}{N}\ (j=0,\ 1,\ \cdots,\ N)$$

为结点的 $f(x)$ 的插值多项式。其系数表达式(2.6.12)利用正交关系(2.6.5)即可导出。

关于系数 a_k 与 A_k 的差别，可以将(2.6.7)代入(2.6.12)，再利用正交关系(2.6.5)并注意到 $\cos\frac{kj\pi}{N}=T_k(x_j)$，$T_{2pN\pm k}(x_j)=T_k(x_j)\ (p=1,\ 2,\ \cdots)$，可得到

$$a_k=\frac{2}{N}\sum_{i=0}^{\infty}{}' A_i\sum_{j=0}^{N}{}'' T_i(x_j)T_k(x_j)=A_k+A_{2N-k}+A_{2N+k}+A_{4N-k}+A_{4N+k}+\cdots \tag{2.6.13}$$

此处 Σ' 表示对 $i=0$ 的项取半。因此，若 N 取得充分大，且系数 A_0，A_1，…是迅速地下降，则 $\bar{S}_N(x)$ 将非常近似于 $S_N(x)$，从而可以用它作为 $f(x)$ 的近似式。

对(2.6.11)式逐项进行积分。应用关系式(2.6.6)，便得到 $f(x)$ 的不定积分的近似表达式

$$\left.\begin{aligned}&\int_{-1}^{x} f(x)\approx I_N(x)\equiv\frac{b_0}{2}+b_1T_1(x)+\cdots+b_NT_N(x)+\cdots\\ &\text{此处}\\ &b_k=\frac{a_{k-1}-a_{k+1}}{2k}\quad(k=1,\ 2,\ \cdots,\ N+1)\\ &b_0=2(b_1-b_2+b_3-\cdots+(-1)^{N-1}b_N+(-1)^Nb_{N+1})\end{aligned}\right\} \tag{2.6.14}$$

其中 $a_k(k=0,\ 1,\ \cdots,\ N)$ 由(2.6.12)给出，但 a_N 取其半，即取 $a_N/2$，而 $a_{N+1}=a_{N+2}=0$，这就是用切氏级数展开积分法的计算公式。

关于积分近似表达式 $I_N(x)=\sum_{k=0}^{N+1}{}' b_kT_k(x)$ 的计算，可以利用递推关系式：

$$\left.\begin{aligned}&c_k=2xc_{k+1}-c_{k+2}+b_k\quad(k=N+1,\ N,\ \cdots,\ 1,\ 0)\\ &c_{N+2}=c_{N+3}=0\end{aligned}\right\} \tag{2.6.15}$$

依次算出数 c_{N+1}，c_N，…，c_0，然后便可得积分值 $I_N(x)$ 为

$$I_N(x)=\frac{1}{2}(c_0-c_2) \tag{2.6.16}$$

这个结果是不难证实的，只要将关系式(2.6.15)代入积分公式(2.6.14)，并利用切比雪夫多项式的递推关系(2.6.3)即可得到。

关于积分公式(2.6.14)的误差，就 $x=1$ 的情形进行讨论，即考虑

$$E_N=\int_{-1}^{1} f(x)\,dx-\int_{-1}^{1}\sum_{k=0}^{N}{}'' a_kT_k(x)\,dx$$

将 $f(x)$ 代入它的切比雪夫展式，经整理后可得

$$E_N=\sum_{k=0}^{N-1}{}'(A_k-a_k)\int_{-1}^{1}T_k(x)\,dx+\left(A_N-\frac{a_N}{2}\right)\int_{-1}^{1}T_N(x)\,dx+\sum_{k=N+1}^{\infty}A_k\int_{-1}^{1}T_k(x)\,dx$$

利用关系式(2.6.6)和(2.6.13)，且假定 N 是偶数以及 A_{3N} 和 A_{3N} 以后的系数可忽略，则得

$$E_N=2\sum_{k=0}^{\frac{N}{2}-1}{}'\frac{A_{2N-2k}+A_{2N+2k}}{(2k)^2-1}-2\sum_{k=1}^{\frac{N}{2}-1}\frac{A_{N+2k}}{(N+2k)^2-1} \tag{2.6.17}$$

因为对于任何一致收敛的切比雪夫级数，其系数 A_k 满足不等式

$$|A_k| \leqslant \frac{c_N}{k}, \quad 当\ k \geqslant N\ 时 \tag{2.6.18}$$

此处 c_N 是与 k 无关的常数。因此有

$$|A_{2N-2k}+A_{2N+2k}| \leqslant \frac{c_N}{2}\left(\frac{1}{N-k}+\frac{1}{N+k}\right)=c_N\frac{N}{N^2-k^2}<\frac{c_N}{N}\left(1+\frac{4k^2}{3N^2}\right) \quad 当\ k<\frac{N}{2}\ 时$$

将此不等式用于式(2.6.17)，则得

$$|E_N|<c_N\left\{\frac{1}{N}+\sum_{k=1}^{\frac{N}{2}-1}\frac{2}{(2k)^2-1}\frac{1}{N}\left(1+\frac{4k^2}{3N^2}\right)\right\}$$

$$=c_N\left\{\frac{2}{N}-\frac{1}{N(N-1)}+\frac{2}{3N^3}\sum_{k=1}^{\frac{N}{2}-1}\frac{4k^2}{4k^2-1}\right\}<\frac{2c_N}{N} \tag{2.6.19}$$

这就是说，当系数 $A_k\ (k\geqslant 3N)$ 可忽略，并且在 A_{N+2} 与 A_{3N} 之间的系数满足不等式 $|A_k|\leqslant\frac{c_N}{k}$，那末就有 $|E_N|<\frac{2c_N}{N}$。因此可以用 $2|A_N|=|a_N|$ 作为 E_N 的粗略估计，为保险起见，可以用相邻三个系数 $2|a_{N-4}|$、$2|a_{N-2}|$、$|a_N|$ 中最大的一个作为误差 $|E_N|$ 的一种估计。

在上面的误差估计中，我们用的条件(2.6.18)是对于坏性态的函数(即其切氏展式的系数是缓慢下降的)来假设的。因此对于好性态的函数(即其切氏展式的系数是快速下降的)，这个估计是过于保守的。对于好性态的函数，不难知道级数

$$\sum_{k=0}^{\infty}{}' A_k\int_{-1}^{1}T_k(x)\,dx=\sum_{k=0}^{\infty}{}' B_kT_k(x)$$

亦是快速收敛的，并且

$$\sum_{k=0}^{N}{}' b_kT_k(x)\approx\sum_{k=0}^{N}{}' B_kT_k(x)$$

因此可以用相邻三个系数 $r^2|b_{N-4}|$、$r|b_{N-2}|$、$|b_N|$ 中最大的一个作为误差 E_N 的一种估计，此处 $0<r<1$，是加权系数，可取 $r=1/8$。

对于不定积分的误差估计也是类似的。所不同的是，此时所有的 $b_k(k=0,\ 1,\ 2,\ \cdots)$ 系数都存在。因此，在慢速收敛的情况，可取 $\max(2|a_{N-2}|,\ 2|a_{N-1}|,\ |a_N|)$ 作为误差估计；在快速收敛的情况，可取 $\max(r^2|b_{N-2}|,\ r|b_{N-1}|,\ |b_N|)$ 作为误差估计。

关于积分公式(2.6.4)中项数 N 的确定，在有了这些误差估计后，我们就可以在计算过程中，按如下的原则自动确定之。

开始取 $N=4$，对于取定的 N 数，在点 $x_j=\cos\frac{j\pi}{N}\ (j=0,\ 1,\ \cdots,\ N)$ 上计算 $f(x)$ 值，然后按公式(2.6.12)和(2.6.14)计算系数 a_k 和 $b_k(k=0,\ 1,\ \cdots,\ N)$。如果有

$$\max(r^2|b_{N-2}|,\ r|b_{N-1}|,\ |b_N|)<\varepsilon \quad (对好性态的函数)$$

或

$$\max(2|a_{N-2}|,\ 2|a_{N-1}|,\ |a_N|)<\varepsilon \quad (对坏性态的函数)$$

此处 ε 是允许误差，则认为此时的 N 已足够了。否则将 N 加倍，即取 $2N$，重复上过程。此时的 $x_j=\cos\frac{j\pi}{2N}(j=0,\ 1,\ \cdots,\ 2N)$ 是在上一次的 x_j 点列上加密，因此前面算出的函数值仍然可以再利用。

用切氏级数展开的积分法，当用于定积分计算时也是一种效率比较高的方法，它的代数

准确度比 N 次要高，并且从计算效果看，其误差通常与高斯型积分公式的误差相近。

上面我们介绍了几种数值积分法。从使用效果上看，这些方法究竟哪种为好？要笼统回答这个问题是比较困难的。因为它们的计算效果依赖于被积函数的性态(如函数的光滑性，以及各阶导数随阶数增高时的量级变化等)。一种积分方法对于不同性态的被积函数，得到的计算效果是不同的。因此对于不同类型的计算对象，有时是这种方法比较好，有时又是另一种方法比较好。例如，在计算积分 $\int_0^1 \frac{2}{2+\sin 10\pi x}dx$ 时，龙勃方法比辛浦生方法优越得多(见下表)。但在计算 $\int_0^1 x^{\frac{1}{2}}dx$ 时，因其被积函数的导数在 $x=0$ 处存在奇异性，故此时辛浦生方法反比龙勃方法的效果为好。

积　　分	真　值	龙 勃 方 法		辛 浦 生 方 法	
		结 点 数	计 算 值	结 点 数	计 算 值
$\int_0^1 \frac{2}{2+\sin 10\pi x}dx$	1.1547005	65	1.1547003	163	1.1546288
		257	1.1547004	883	1.1547002
$\int_0^1 x^{\frac{1}{2}}dx$	0.66666667	65	0.66653263	55	0.66665866
		4097	0.66666633	199	0.66666655

因此，在作方法选择时，需针对具体情况作具体的分析。

梯形和辛浦生方法是两种低精度的方法，但对于光滑性差的被积函数，有时比使用其他高精度方法的效果要好一些。尤其是梯形方法对于周期的被积函数具有特殊好的效果(见§2.7)。

下面我们主要对其他三种方法，即高斯-勒让德积分法、龙勃方法以及用切氏级数展开的积分法，在准确性、稳定性以及使用方便性等几方面作一些粗略的分析和比较。

(1) 方法的准确度　即在一定的工作量(计算被积函数的数目)下，由方法产生的积分近似所具有的准确度。

关于方法准确度的一种度量是考察其公式的截断误差。对于具有 $N=2^k+1$(k 是正整数)形状的结点数，三个公式的截断误差或其估式分别是：

高斯-勒让德公式的截断误差(见公式(2.5.17))：

$$E_N=\int_{-1}^1 f(x)-\sum_{i=1}^N c_i f(x_i)=\frac{4^{k+1}(2^k+1)[2^k!]^4}{(2^{k+1}+3)[(2^{k+1}+1)!]^3}f^{(2^{k+1}+2)}(\xi)\quad -1<\xi<1$$

龙勃方法(取主对角线序列时)的截断误差(参见资料[7]、[8])

$$E_N=T_{k0}-\int_{-1}^1 f(x)dx=\frac{B_{2k+2}}{2^{k^2-k-3}(2k+2)!}f^{(2k+2)}(\xi)\quad -1<\xi<1$$

式中 B_{2k+2} 是贝努利常数。

用切氏级数展开积分法的截断误差估式(参见资料[9])有

$$|E_N|=\left|\int_{-1}^1 f(x)dx-\sum_{i=0}^{2^k}{}'' a_iT_i(x)\right|\leqslant\frac{2}{4^k(2^k+1)!}\max_{-1<x<1}|f^{(2^k+1)}(x)|$$

如果可以不考虑在截断误差中高阶导数部分的作用，那末就可以单从其系数部分的大小来进行比较。在[8]中，对高斯-勒让德和龙勃方法作了这方面的比较，并发现高斯-勒让德积分公式在这点上是优于龙勃方法的。例如，对 $N=65$，龙勃方法的截断误差的系数部分

为 0.8693×10^{-8}，而高斯-勒让德公式在 $N=64$ 时其截断误差的系数部分已为 0.3×10^{-18}。在文中还从一个例子的计算中表明了，采用10点的高斯-勒让德公式和65点的龙勃方法的精度差不多。

用切氏级数展开的积分法，从其误差估式看，其系数部分是优于龙勃方法的。

自然这种比较并不适合所有的情况，因为有时高阶导数部分也会起很大的影响。这三个误差公式具有不同阶的高阶导数，其中高斯-勒让德公式的阶数最高。因此当遇导数量值随阶数增高而增长的情况时，它在这方面又最不利。此时它们的误差就可能与上述结论不符了。

在[11]中，对此三种方法作了很多试算例子的比较。从这些例子来看，高斯-勒让德公式的精度最高；用切氏级数展开的积分法与高斯-勒让德公式的精度相近(对好性态和坏性态的函数均如此)；对好性态的函数，用切氏级数展开的积分法优于龙勃方法。

(2) 方法的稳定性　因为我们的计算是有尽数位的计算，计算结果的准确度除了与积分公式准确度有关外，还与方法的稳定性即舍入误差的影响有关。如果我们把积分方法的算式写成如下形式的积分公式

$$\int_a^b f(x)\,dx\approx\sum_{i=1}^{N}W_i f(x_i)$$

式中 x_i 是结点；W_i 是积分系数。如果这些系数 W_i 都具有相同符号并且量级相近，那末这种积分公式的稳定性就比较好。在龙勃方法中这些系数都是非负的，并且可以证明，其相互间在数量上的差别不超过三倍。高斯-勒让德积分公式的系数都是正的，但其量级上的差别比较大(见本章的附表一)。用切氏级数展开的积分法，它的积分系数亦都是正的，且对任何 N 均有 $0<W_i<2(i=1,\ 2,\ \cdots,\ N)$(参考[10])，但其量级上的差别还是存在的，因此从稳定性角度看，龙勃方法比较好。

(3) 方法使用的方便性　龙勃方法的算法比较简单，并且当加密点列作提高积分近似度的计算时，前面计算过的函数值可全部沿用，而这部分工作占了总工作量的一半。用切氏级数展开积分法亦具有此种性质，但其计算方法比龙勃方法稍复杂一些。这两种方法都能方便地得到积分的近似值序列，并存在比较简单的误差估算法。除此，龙勃方法还有另一方便性，因它同时得到若干个积分序列，特别是得到梯形和辛浦生序列。如果在作收敛性控制时同时检验主对角线序列和梯形、辛浦生序列，那末对于不同性态的函数就可以用其中最快的收敛序列来逼近积分值。

高斯-勒让德公式在这些方面刚好是它的弱点。因为它的结点是不规则的，当结点数 N 增加时，前面计算的函数值完全不能被后面利用。并且在使用中，需预先存入一串不同 N 值的结点值和积分系数表，或者需通过比较复杂的计算来形成它们。因此在逐步逼近积分(即逐步增加结点数 N)的计算过程中不如前两种方法方便。

§2.7 在离散点上给出函数的积分，平均抛物插值法

前面介绍的几种积分法，其积分坐标点的选取不是随意的，它或者要求取等距结点，或者要求取某些特定的结点。因此，如果被积函数不是用公式给出的函数，而是用实验或测量得到的一串数据，即我们只知道被积函数在若干离散点 $x_1,\ x_2,\ \cdots,\ x_n$ 上的数值，那末，这些方法就无法使用了。因为这些实验(或测量)点 $x_i(i=1,\ 2,\ \cdots,\ n)$ 不一定是等距的，也不会

恰好落在那些特定点上。

对于这种在离散点上给出量值的函数，可以采用分段插值法来计算它的积分。下面介绍其中一种，这个方法是在被积函数的抛物插值和平均的平滑处理基础上构造的，现将方法叙述如下：

设被积函数$f(x)$是一在下列离散点上给出量值的函数。

$$\begin{array}{llllll} x: & x_0 & x_1 & \cdots & x_{n-1} & x_n \\ f: & f_0 & f_1 & \cdots & f_{n-1} & f_n \end{array}$$

此处坐标点x_i有

$$a=x_0<x_1<\cdots<x_{n-1}<x_n=b$$

计算积分$I=\int_a^b f(x)dx$的平均抛物插值法，就是在每个小区段$[x_i, x_{i+1}]$上，通过点(x_{i-1}, f_{i-1})、(x_i, f_i)、(x_{i+1}, f_{i+1})作二次抛物线

$$P(x; x_{i-1}, x_i, x_{i+1})=a_i x^2+b_i x+c_i$$

并通过点(x_i, f_i)、(x_{i+1}, f_{i+1})、(x_{i+2}, f_{i+2})作二次抛物线

$$P(x; x_i, x_{i+1}, x_{i+2})=a_{i+1}x^2+b_{i+1}x+c_{i+1}$$

取其平均值，得

$$P_i(x)=\frac{1}{2}[P(x; x_{i-1}, x_i, x_{i+1})+P(x; x_i, x_{i+1}, x_{i+2})]$$

以此二次函数$P_i(x)$来近似于函数$f(x)$，并以$P_i(x)$在小区间$[x_i, x_{i+1}]$上的积分来近似于$f(x)$的积分，即

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \int_{x_i}^{x_{i+1}} P_i(x)dx=\frac{1}{2}\int_{x_i}^{x_{i+1}}[P(x; x_{i-1}, x_i, x_{i+1})+P(x; x_i, x_{i+1}, x_{i+2})]dx \tag{2.7.1}$$

利用拉格朗日公式，且令

$$\left.\begin{aligned} h_i&=x_{i+1}-x_i \\ \delta_i&=h_i/h_{i-1} \\ \lambda_i&=h_i/h_{i+1} \end{aligned}\right\} \tag{2.7.2}$$

可写出$P(x; x_{i-1}, x_i, x_{i+1})$，$P(x; x_i, x_{i+1}, x_{i+2})$的表达式

$$P(x; x_{i-1}, x_i, x_{i+1})=L_i\left(\frac{x-x_i}{h_i}\right)^2-(L_i+f_i-f_{i+1})\left(\frac{x-x_i}{h_i}\right)+f_i \tag{2.7.3}$$

式中

$$L_i=\frac{\delta_i^2}{1+\delta_i}f_{i-1}-\delta_i f_i+\frac{\delta_i}{1+\delta_i}f_{i+1} \tag{2.7.4}$$

$$P(x; x_i, x_{i+1}, x_{i+2})=R_i\left(\frac{x-x_i}{h_i}\right)^2-(R_i+f_i-f_{i+1})\left(\frac{x-x_i}{h_i}\right)+f_i \tag{2.7.5}$$

式中

$$R_i=\frac{\lambda_i}{1+\lambda_i}f_i-\lambda_i f_{i+1}+\frac{\lambda_i^2}{1+\lambda_i}f_{i+2} \tag{2.7.6}$$

不难写出它们在区间$[x_i, x_{i+1}]$上的积分

$$\int_{x_i}^{x_{i+1}} P(x; x_{i-1}, x_i, x_{i+1})dx=\frac{h_i}{6}(3f_i+3f_{i+1}-L_i)$$

$$\int_{x_i}^{x_{i+1}} P(x; x_i, x_{i+1}, x_{i+2})dx = \frac{h_i}{6}(3f_i + 3f_{i+1} - R_i)$$

于是得

$$\int_{x_i}^{x_{i+1}} P_i(x)dx = \frac{h}{6}\left(3f_i + 3f_{i+1} - \frac{L_i + R_i}{2}\right) \quad (i=1, 2, \cdots, n-2) \tag{2.7.7}$$

对于第一和最后一个小区间上的积分，不作平滑处理，而用

$$\int_{x_0}^{x_1} P(x; x_0, x_1, x_2)dx = \frac{h_0}{6}(3f_0 + 3f_1 - R_0) \tag{2.7.8}$$

$$\int_{x_{n-1}}^{x_n} P(x; x_{n-2}, x_{n-1}, x_n)dx = \frac{h_{n-1}}{6}(3f_{n-1} + 3f_n - L_{n-1}) \tag{2.7.9}$$

将这些小区间上的积分相加起来，便得积分近似式

$$\int_a^b f(x)dx \approx I_n = \frac{1}{6}\sum_{i=0}^{n-1} h_i\left(3f_i + 3f_{i+1} - \frac{L_i + R_i}{2}\right) \tag{2.7.10}$$

此处 L_i, R_i 由式(2.7.4)和(2.7.6)以及(2.7.2)给出，但其中令 $L_0=R_0$, $R_{n-1}=L_{n-1}$。

若积分下限 a 不是落在给定的坐标点 x_0 上，则有二种情形:

(1) 若 $x_i<a<x_{i+1}$, 而 $1\leqslant i\leqslant n-2$, 则在小区间 $[a, x_{i+1}]$ 上的积分采用近似式

$$\begin{aligned}\int_a^{x_{i+1}} f(x)dx &\approx \frac{1}{2}\int_a^{x_{i+1}}[P(x; x_{i-1}, x_i, x_{i+1}) + P(x; x_i, x_{i+1}, x_{i+2})]dx \\ &= \frac{h_i}{6}\left(3f_i + 3f_{i+1} - \frac{L_i + R_i}{2}\right) \\ &\quad - \frac{a - x_i}{6}\left[(L_i + R_i)\left(\frac{a - x_i}{h_i}\right)^2 - 3\left(\frac{L_i + R_i}{2} + f_i - f_{i+1}\right)\left(\frac{a - x_i}{h_i}\right) + 6f_i\right]\end{aligned} \tag{2.7.11}$$

(2) 若 $x_0<a<x_1$ 或 $a<x_0$, 则在小区间 $[a, x_1]$ 上的积分采用近似式

$$\begin{aligned}\int_a^{x_1} f(x)dx &\approx \int_a^{x_1} P(x; x_0, x_1, x_2)dx \\ &= \frac{h_0}{6}(3f_0 + 3f_1 - R_0) \\ &\quad - \frac{a - x_0}{6}\left[2R_0\left(\frac{a - x_0}{h_0}\right)^2 - 3(R_0 + f_0 - f_1)\left(\frac{a - x_0}{h_0}\right) + 6f_0\right]\end{aligned} \tag{2.7.12}$$

若令 $L_0=R_0$, 则可将式(2.7.12)统一成式(2.7.11)。

类似地，若积分上限 b 不落在坐标点 x_n 上，这时亦分两种情形:

(1) $x_i<b<x_{i+1}$, 而 $1\leqslant i\leqslant n-2$, 则在小区间 $[x_i, b]$ 上的积分采用近似式

$$\begin{aligned}\int_{x_i}^b f(x)dx &\approx \frac{1}{2}\int_{x_i}^b [P(x; x_{i-1}, x_i, x_{i+1}) + P(x; x_i, x_{i+1}, x_{i+2})]dx \\ &= \frac{b - x_i}{6}\left[(L_i + R_i)\left(\frac{b - x_i}{h_i}\right)^2 - 3\left(\frac{L_i + R_i}{2} + f_i - f_{i+1}\right)\left(\frac{b - x_i}{h_i}\right) + 6f_i\right]\end{aligned} \tag{2.7.13}$$

(2) $x_{n-1}<b<x_n$, 或 $x_n<b$, 则在小区间 $[x_{n-1}, b]$ 上的积分采用近似式

$$\begin{aligned}\int_{x_{n-1}}^b f(x)dx &\approx \int_{x_{n-1}}^b P(x; x_{n-2}, x_{n-1}, x_n)dx \\ &= \frac{b - x_{n-1}}{6}\left[2L_{n-1}\left(\frac{b - x_{n-1}}{h_{n-1}}\right)^2 - 3(L_{n-1} + f_{n-1} - f_n)\left(\frac{b - x_{n-1}}{h_{n-1}}\right) + 6f_{n-1}\right]\end{aligned} \tag{2.7.14}$$

若令 $R_{n-1}=L_{n-1}$, 则式(2.7.14)可统一成式(2.7.13)。

§2.8 周期函数的积分

设 $f(x)$ 是一周期为 l 的连续函数,要求计算 $f(x)$ 在其一个周期上的积分。由于函数的周期性,其积分值 I 与积分区间位置的选择是无关的。任何长度为 l 的区间 $[a, b]$ 均可取为积分区间,这时有

$$I=\int_a^b f(x)dx \quad (b-a=l)$$

这也就是说,对于积分区间的任何位移,其积分值是不变的。

鉴于周期函数的这一性质,自然希望选择的数值积分公式亦具有这种位移不变性,即当积分区间作任何位移时,积分近似式不变。下面来考察一下这样的积分公式的特点。

设我们取等距结点的积分公式,

$$\begin{aligned}I&=\int_a^b f(x)dx\approx W_0f(a)+W_1f(a+h)+\cdots+W_{n-1}f(a+(n-1)h)+W_nf(b)\\&=(W_0+W_n)f(a)+W_1f(a+h)+\cdots+W_{n-1}f(a+(n-1)h)\end{aligned}\tag{2.8.1}$$

式中 $b-a=l$; $h=l/n$; $W_0, W_1, \cdots, W_n$ 是积分系数。在等式中应用了函数的周期性。若将积分区间右移一h, 则得

$$\begin{aligned}I&=\int_{a+h}^{b+h} f(x)dx\approx W_0f(a+h)+\cdots+W_{n-2}f(a+(n-1)h)+W_{n-1}f(b)+W_nf(b+h)\\&=(W_0+W_n)f(a+h)+\cdots+W_{n-2}f(a+(n-1)h)+W_{n-1}f(a)\end{aligned}\tag{2.8.2}$$

近似式(2.8.1)和(2.8.2)左端的积分是相等的。若要求它们右端的积分近似亦相同,则比较上两式即可得

$$W_0+W_n=W_{n-1}, \quad W_1=W_0+W_n, \quad W_2=W_1, \cdots, \quad W_{n-1}=W_{n-2}$$

由此知,当积分系数具有关系

$$W_0+W_n=W_1=W_2=\cdots=W_{n-2}=W_{n-1}\tag{2.8.3}$$

时,积分公式(2.8.1)便具有上述的位移不变性。

我们知道梯形积分公式的积分系数是满足条件(2.8.3)的,因此,应用梯形公式来计算周期函数在其一个周期上的积分,当积分区间作任何位移时,都将给出相同的积分近似式,且有

$$\begin{aligned}I&=\int_a^b f(x)dx\approx h\left[\frac{1}{2}f(a)+f(a+h)+\cdots+f(a+(n-1)h)+\frac{1}{2}f(b)\right]\\&=h[f(a)+f(a+h)+\cdots+f(a+(n-1)h)]\end{aligned}$$

实际上化成了简单的矩形公式。

梯形公式用于周期函数,不仅具有这种合理性和简单性,而且在精确度上效果也格外好,其精确性远比式(2.1.11)给出的估计要好得多。事实上可以证明(可见[1]):

设 $f(x)$ 是具有以下性质的函数:

(1) $f(x)$ 在 $[a, b]$ 上存在 $2k+1$ 阶连续导数,且在 $[a, b]$ 上有 $|f^{(2k+1)}(x)|\leqslant M$;

(2) 若 $f'(a)=f'(b), f'''(a)=f'''(b), \cdots, f^{(2k-1)}(a)=f^{(2k-1)}(b)$, 则对积分 $\int_a^b f(x)\,dx$ 应用梯形公式

$$T_n=h\left[\frac{1}{2}f(a)+f(a+h)+\cdots+f(a+(n-1)h)+\frac{1}{2}f(b)\right] \qquad h=\frac{b-a}{n}$$

将有误差估计式

$$\left|\int_a^b f(x)dx-T_n\right|\leqslant\frac{c}{n^{2k+1}} \tag{2.8.4}$$

式中

$$c=2M\frac{(b-a)^{2k+2}}{(2\pi)^{2k+1}}\sum_{j=1}^{\infty}j^{-(2k+1)} \tag{2.8.5}$$

因此当周期函数足够光滑时，即具有足够高阶的导数时，将有非常高阶的误差估式。

§2.9 奇异积分，不连续的被积函数

前面介绍的几种数值积分法，其误差估计都是在被积函数及其若干阶导数为连续的假定下作出的。因此，如果被积函数不满足这些条件，那末在使用这些积分公式时，其误差将有可能比估计式所指出的要大得多；在某些情况下，甚至根本不能作为积分近似。因此一般需作适当的处理，方能获得满意的结果。

2.9.1 存在有限个间断点的有界的被积函数

如果有界的被积函数 $f(x)$ 在积分区间 $[a, b]$ 上存在若干个间断点，譬如说存在一个间断点，位于 $x=c(a<c<b)$ 上，那末我们可以用间断点分割积分区间，写出积分

$$\int_a^b f(x)dx=\int_a^c f(x)dx+\int_c^b f(x)dx \tag{2.9.1}$$

如果 $f(x)$ 在区间 $[a, c]$, $[c, b]$ 上存在足够高阶的连续导数，则右端的两个积分可以用任何一种积分法来近似到所要的精度。

如果被积函数是连续的，但其某个低阶导数有一不连续点，则亦可用类似的办法处理。

2.9.2 无界的被积函数

设积分 $\int_a^b f(x)dx$，当 $x\to a$ 时 $f(x)\to\infty$。如果极限 $\lim\limits_{\varepsilon\to 0}\int_{a+\varepsilon}^b f(x)dx$ 存在，则称此积分存在，且定义为

$$\int_a^b f(x)dx=\lim_{\varepsilon\to 0}\int_{a+\varepsilon}^b f(x)dx \tag{2.9.2}$$

对于 $f(x)$ 在积分上限处为无界的情形，其定义和处理方法都是类似的。因此我们仅就一种情形进行讨论。

这类奇异积分，有时可以采用变量的变换，或者采用分部积分法，将它化为正常积分。例如：

例 1　积分

$$\int_0^1 x^{-\frac{1}{n}}f(x)dx, \quad n\geqslant 2$$

其中 $f(x)$ 在 $[0, 1]$ 上连续，则可施以 $t^n=x$ 的变量变换，将积分化为正常积分

$$n\int_0^1 f(t^n)t^{n-2}dt$$

例2 积分

$$\int_0^1 \frac{\cos x}{\sqrt{x}} dx$$

则可以通过分部积分后，得

$$\int_0^1 \frac{\cos x}{\sqrt{x}} dx = 2\sqrt{x}\cos x \Big|_0^1 - 2\int_0^1 \sqrt{x}\sin x dx$$

这样就化为对函数$\sqrt{x}\sin x$的积分。它在$[0, 1]$上已无奇异点(但有一无界的导数的点)。

如果不能通过类似的数学演化工作来消除它的奇异性，那末便需在积分计算中进行适当的处理，或者采用特殊的积分方法。下面提供几种这样的方法。

一、区间的截去

适当地选取小数δ，使得在小区间$[a, a+\delta]$上的积分值处在允许的误差范围之内，即有

$$\left|\int_a^{a+\delta} f(x)dx\right| < \varepsilon$$

那末就可计算正常积分

$$\int_{a+\delta}^b f(x)dx$$

来近似在全区间上的积分。

例如，计算积分

$$\int_0^1 \frac{g(x)}{x^{\frac{1}{2}}+x^{\frac{1}{3}}} dx$$

其中$g(x)$在$[0, 1]$上连续，且满足$|g(x)| \leqslant 1$。

因为在$[0, 1]$上，$x^{\frac{1}{2}} \leqslant x^{\frac{1}{3}}$，则有

$$\left|\frac{g(x)}{x^{\frac{1}{2}}+x^{\frac{1}{3}}}\right| \leqslant \frac{1}{2x^{\frac{1}{2}}}$$

因此

$$\left|\int_0^\delta \frac{g(x)}{x^{\frac{1}{2}}+x^{\frac{1}{3}}} dx\right| \leqslant \frac{1}{2}\int_0^\delta \frac{dx}{x^{\frac{1}{2}}} = \delta^{\frac{1}{2}}$$

由此得到，对于10^{-3}的精度要求，可取$\delta \leqslant 10^{-6}$，由计算在$[\delta, 1]$上的正常积分来得到所要的积分值。

二、分项法

为易于说明起见，假定积分具有如下的形式：

$$\int_a^b (x-a)^\alpha f(x)dx \quad (-1 < \alpha \leqslant 0)$$

在$x=a$处，被积函数存在一奇异点，但$f(x)$是一充分光滑的函数，假定它存在$n+1$阶的连续导数，则可以写为

$$f(x) = g(x) + [f(x) - g(x)]$$

式中

$$g(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

而积分

$$\int_a^b (x-a)^\alpha f(x)\,dx=\int_a^b (x-a)^\alpha g(x)\,dx+\int_a^b (x-a)^\alpha[f(x)-g(x)]\,dx$$

右端的第一个积分，其被积函数

$$(x-a)^\alpha g(x)=f(a)(x-a)^\alpha+f'(a)(x-a)^{1+\alpha}+\cdots+\frac{f^{(n)}(a)}{n!}(x-a)^{n+\alpha}$$

可以逐项求出其积分值；右端的第二个积分，其被积函数在$[a,\ b]$上已无奇异性，并且相当光滑，因此可以利用任何一种求积公式来得到它的近似值。

例如　计算

$$\int_0^1 \frac{\cos x}{\sqrt{x}}\,dx$$

因为

$$\cos x=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots+(-1)^n\frac{x^{2n}}{(2n)!}+\cdots$$

我们可取其若干项，譬如取二项，可得积分

$$\int_0^1 \frac{\cos x}{\sqrt{x}}\,dx=2-\frac{1}{5}+\int_0^1 \frac{\cos x-1+\frac{1}{2}x^2}{\sqrt{x}}\,dx$$

右端积分的被积函数在$[0,\ 1]$上存在连续的三阶导数。如要使函数更光滑些，可以在$\cos x$的展式中再多取几项。

三、插值型积分公式

将奇异的被积函数$F(x)$分解成两个函数的乘积

$$F(x)=W(x)f(x)$$

使奇异性全部集中在函数$W(x)$中，而$f(x)$是一足够光滑的函数。

在$[a,\ b]$内适当选择一串坐标点

$$x_0<x_1<x_2<\cdots<x_n$$

构造积分公式

$$\int_a^b W(x)f(x)\,dx=\sum_{i=0}^n W_i f(x_i)$$

使得对任何n次多项式是准确的，亦即有

$$\int_a^b W(x)x^k=\sum_{i=0}^n W_i x_i^k\quad(k=0,\ 1,\ \cdots,\ n)$$

假设这$n+1$个方程左端的积分均存在，并且可以具体算出它们的量值，那末就可以从此线性方程组定出$n+1$个W_i值，而得到计算积分的近似公式。

例如　计算积分

$$\int_0^1 x^{-\frac{1}{2}}f(x)\,dx$$

其中$f(x)$在$[0,\ 1]$上连续。我们取3个坐标点：$x_0=\frac{1}{3}$，$x_1=\frac{2}{3}$，$x_3=1$，得线性方程组

$$W_1+W_2+W_3=\int_0^1 x^{-\frac{1}{2}}\,dx=2$$

$$\frac{W_1}{3}+\frac{2W_2}{3}+W_3=\int_0^1 x^{-\frac{1}{2}}\,x\,dx=\frac{2}{3}$$

$$\frac{W_1}{9}+\frac{4W_2}{9}+W_3=\int_0^1 x^{-\frac{1}{2}}x^2dx=\frac{2}{5}$$

由此导出积分公式

$$\int_0^1 x^{-\frac{1}{2}}f(x)dx\approx\frac{14}{5}f\left(\frac{1}{3}\right)-\frac{8}{5}f\left(\frac{2}{3}\right)+\frac{4}{5}f(1)$$

四、高斯型积分公式

对于奇异积分常可应用高斯型求积公式来解决，只要奇异的被积函数 $F(x)$ 能分解成

$$F(x)=W(x)f(x)$$

其中 $W(x)$ 是一固定的非负的权函数，它在积分区间 $[a, b]$ 上存在一个或几个奇异点，并且积分 $\int_a^b W(x)x^kdx$ $(k=0, 1, \cdots, n)$ 存在。而 $f(x)$ 是一足够光滑的函数。那末，可以按照 §2.3 介绍的方法来确定积分公式的结点和系数。下面列举 $W(x)$ 的几种情形：

(i) $W(x)=(1-x^2)^{-\frac{1}{2}}$，相应得高斯-切比雪夫积分公式。详见§2.3。

(ii) $W(x)=(1-x^2)^{\frac{1}{2}}$，相应得积分公式

$$\int_{-1}^1\sqrt{1-x^2}f(x)dx=\sum_{k=1}^n W_kf(x_k)$$

此处

$$x_k=\cos\frac{k}{n+1}\pi,\quad W_k=\frac{\pi}{n+1}\sin^2\frac{k}{n+1}\pi$$

截断误差为

$$E_n=\frac{\pi}{(2n)!2^{2n+1}}f^{(2n)}(\xi)\quad -1<\xi<1$$

(iii) $W(x)=x^{\frac{1}{2}}(1-x)^{-\frac{1}{2}}$，相应得积分公式

$$\int_0^1\sqrt{\frac{x}{1-x}}f(x)dx=\sum_{k=1}^n W_kf(x_k)$$

此处

$$x_k=\cos^2\frac{2k-1}{2n+1}\frac{\pi}{2},\quad W_k=\frac{2\pi}{2n+1}x_k$$

截断误差为

$$E_n=\frac{\pi}{(2n)!2^{4n+1}}f^{(2n)}(\xi)\quad 0<\xi<1$$

(iv) $W(x)=(1-x)^{\frac{1}{2}}$，相应得积分公式

$$\int_0^1\sqrt{1-x}f(x)dx=\sum_{k=1}^n W_kf(x_k)$$

此处

$x_k=1-z_k^2$（z_k 是 $2n+1$ 次勒让德多项式 $L_{2n+1}(x)$ 的第 k 个正零点）。

$$W_k=2z_k^2W_k^{(2n+1)}$$

式中

$W_k^{(2n+1)}=\dfrac{2}{(1-z_k^2)[L'_{2n+1}(z_k)]^2}$ 是 $2n+1$ 点高斯-勒让德公式中相应于结点 z_k 的系数。

截断误差为

$$E_n=\frac{2^{4n+3}[(2n+1)!]^4}{(2n)!(4n+3)[(4n+2)!]^2}f^{(2n)}(\xi)\quad 0<\xi<1$$

(v) $W(x)=(1-x)^{-\frac{1}{2}}$, 相应得积分公式

$$\int_0^1\frac{f(x)}{\sqrt{1-x}}dx=\sum_{k=1}^{n}W_kf(x_k)$$

此处

$$x_k=1-z_k^2\quad(z_k\text{ 是 }2n\text{ 次勒让德多项式 }L_{2n}(x)\text{ 的第 }k\text{ 个正零点})$$

$$W_k=2W_k^{(2n)}$$

式中

$$W_k^{(2n)}=\frac{2}{(1-z_k^2)[L'_{2n}(z_k)]^2}$$ 是 $2n$ 点高斯-勒让德公式中相应于结点 z_k 的系数。

截断误差为

$$E_n=\frac{2^{4n+1}}{4n+1}\frac{[(2n)!]^3}{[(4n)!]^3}f^{(2n)}(\xi)\quad 0<\xi<1$$

五、柯西主值的数值计算

考虑积分 $\int_a^b f(x)dx$, 假定 $f(x)$ 在 $x=c\,(a<c<b)$ 的邻域内无界。此时的积分定义有两种:

(i) 如果积分

$\int_a^c f(x)dx$, $\int_c^b f(x)dx$ 存在, 即如果极限 $\lim\limits_{\varepsilon\to 0}\int_a^{c-\varepsilon}f(x)dx$, $\lim\limits_{\delta\to 0}\int_{c+\delta}^b f(x)dx$ 存在, 则定义

$$\int_a^b f(x)dx=\int_a^c f(x)dx+\int_c^b f(x)dx$$

(ii) 如果极限

$$\lim_{\varepsilon\to 0}\left[\int_a^{c-\varepsilon}f(x)dx+\int_{c+\varepsilon}^b f(x)dx\right]$$

存在, 则此极限称为定积分的柯西主值, 且记为 $P\int_a^b f(x)dx$。

当两种极限都存在时, 这两种定义是一致的。但在 i) 之极限不存在时, ii) 之极限亦有可能存在。

关于柯西主值的计算, 可以利用下面的分解办法, 将它化为一个正常积分或者化为在积分下限处的奇异积分。

不妨假定 $c=0$, 且假定计算积分

$$\int_{-a}^a f(x)dx$$

则可作函数

$$g(x)=\frac{1}{2}[f(x)-f(-x)],\qquad h(x)=\frac{1}{2}[f(x)+f(-x)]$$

显见 $g(x)$ 是一奇函数, 即 $g(-x)=-g(x)$; $h(x)$ 是一偶函数, 即 $h(-x)=h(x)$, 而

$$f(x)=g(x)+h(x)$$

因此

$$\int_{-a}^{-\varepsilon}f(x)dx+\int_\varepsilon^a f(x)dx=\int_{-a}^{-\varepsilon}g(x)dx+\int_\varepsilon^a g(x)dx+\int_{-a}^{-\varepsilon}h(x)dx+\int_\varepsilon^a h(x)dx=2\int_\varepsilon^a h(x)dx$$

于是得柯西主值

$$P\int_{-a}^{a} f(x)\,dx = 2\lim_{\varepsilon\to 0}\int_{\varepsilon}^{a} h(x)\,dx = 2\int_{0}^{a} h(x)\,dx = \int_{0}^{a}[f(x)+f(-x)]\,dx$$

$h(x)$在某些情况下可能已无奇异性，这时就化为一正常积分的计算；在一般情况下，$h(x)$在$x=0$处存在一奇异点，这就化为在积分下限处的奇异积分。

§2.10 在无穷区间上的积分

在实际工作中，我们也经常遇见在无穷区间上的积分，即需计算积分

$$\int_{a}^{\infty} f(x)\,dx \text{ 或 } \int_{-\infty}^{b} f(x)\,dx \text{ 或 } \int_{-\infty}^{\infty} f(x)\,dx$$

这样的积分是由正常积分的极限来定义的。

对单无穷区间上的积分，如果下式右端的极限存在，则定义为

$$\int_{a}^{\infty} f(x)\,dx = \lim_{R\to\infty}\int_{a}^{R} f(x)\,dx$$

$$\int_{-\infty}^{b} f(x)\,dx = \lim_{R\to\infty}\int_{-R}^{b} f(x)\,dx$$

对于$[-\infty, \infty]$上的积分，有两种定义：

(1) $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx$

(2) 积分的柯西主值 $P\int_{-\infty}^{\infty} f(x)\,dx = \lim_{R\to\infty}\int_{-R}^{R} f(x)\,dx$

当两种极限都存在时，这两种定义是一致的。但在(1)之极限不存在时，(2)之极限有可能存在。

下面就积分

$$\int_{a}^{\infty} f(x)\,dx$$

提供几种处理方法。对于$[-\infty, b]$上的积分，其处理是类似的。

一、变量的改变

在某些情况下，可以通过变量变换，将无穷区间上的积分变成一有限区间上的积分。

例如　令$t=e^{-x}$，将区间$0\leqslant x\leqslant\infty$变到区间$0\leqslant x\leqslant 1$，而积分变成

$$\int_{0}^{\infty} f(x)\,dx = \int_{0}^{1}\frac{f(-\ln t)}{t}\,dt$$

或

$$\int_{0}^{\infty} e^{-x} f(x)\,dx = \int_{0}^{1} f\left(\ln\frac{1}{t}\right)dt$$

二、无穷区间的截去

适当地选择一数$R>a$，使得积分$\int_{R}^{\infty} f(x)\,dx$的量值处在允许的误差范围之内，即有

$$\left|\int_{R}^{\infty} f(x)\,dx\right| < \varepsilon$$

则可以计算有限区间上的积分

$$\int_{a}^{R} f(x)\,dx \approx \int_{a}^{\infty} f(x)\,dx$$

来近似无穷区间上的积分。

例如 计算

$$\int_0^\infty e^{-x^2}dx$$

因为对 $x\geqslant k$ 有 $x^2\geqslant kx$, 故有估式

$$\int_k^\infty e^{-x^2}dx\leqslant\int_k^\infty e^{-kx}dx=\frac{e^{-k^2}}{k}$$

对于 $k=4$, 有 $e^{-k^2}/k\approx10^{-8}$, 因此对于 10^{-7} 的精度要求, 计算积分 $\int_0^4 e^{-x^2}dx$ 即可满足。

三、高斯型积分公式

高斯型积分公式常可有效地用于计算无穷区间上的积分。可以建立积分公式

$$\int_0^\infty W(x)f(x)dx\approx\sum_{k=1}^n W_k f(x_k)$$

$$\int_{-\infty}^\infty W(x)f(x)dx\approx\sum_{k=1}^n W_k f(x_k)$$

使对任何 $2n-1$ 次多项式是正确的。建立公式的方法可见 §2.4。在实际中经常使用的有以下几个公式:

(i) 拉盖尔(Laguerre)公式

相应于权函数 $W(x)=e^{-x}$, 其公式为

$$\int_0^\infty e^{-x}f(x)dx\approx\sum_{k=1}^n W_k f(x_k)$$

此处坐标 $x_k(k=1, \cdots, n)$ 是拉盖尔多项式

$$L_n(x)=(-1)^n e^x\frac{d^n}{dx^n}(x^n e^{-x})=x^n+\cdots$$

的 n 个零点, 而系数

$$W_k=\frac{(n!)^2 x_k}{[L_{n+1}(x_k)]^2}$$

公式的截断误差是

$$E_n=\frac{(n!)^2}{(2n)!}f^{(2n)}(\xi)\quad 0<\xi<\infty$$

计算拉盖尔多项式, 可以采用递推公式

$$L_{n+2}(x)=[x-(2n+3)]L_{n+1}(x)-(n+1)^2L_n(x)$$

$$L_0(x)=1$$

$$L_1(x)=x-1$$

(ii) 广义拉盖尔公式

相应于权函数 $W(x)=x^\alpha e^{-x}(\alpha>-1)$, 其公式为

$$\int_0^\infty x^\alpha e^{-x}f(x)dx\approx\sum_{k=1}^n W_k f(x_k)$$

此处坐标 $x_k(k=1, \cdots, n)$ 是广义拉盖尔多项式

$$L_n^{(\alpha)}(x)=(-1)^n x^{-\alpha}e^x\frac{d^n}{dx^n}(x^{\alpha+n}e^{-x})=x^n+\cdots$$

的 n 个零点, 系数

$$W_k=\frac{n!\Gamma(n+\alpha+1)x_k}{[L_{n+1}^{(\alpha)}(x_k)]^2}$$

公式的截断误差是

$$E_n=\frac{n!\Gamma(n+\alpha+1)}{(2n)!}f^{(2n)}(\xi)\quad 0<\xi<\infty$$

计算广义拉盖尔多项式，可以采用递推公式

$$L_{n+2}^{(\alpha)}(x)=[x-\alpha-(2n+3)]L_{n+1}^{(\alpha)}(x)-(n+1)(n+\alpha+1)L_n^{(\alpha)}(x)$$

$$L_0^{(\alpha)}(x)=1$$

$$L_1^{(\alpha)}(x)=x-(\alpha+1)$$

(iii) 埃尔米特(Hermite)公式

相应于权函数 $W(x)=e^{-x^2}$，其公式为

$$\int_{-\infty}^{\infty}e^{-x^2}f(x)dx\approx\sum_{k=1}^{n}W_kf(x_k)$$

此处坐标 $x_k(k=1,\ \cdots,\ n)$ 是埃尔米特多项式

$$H_n(x)=(-1)^ne^{x^2}\frac{d^n}{dx^n}(e^{-x^2})=2^nx^n+\cdots$$

的 n 个零点，系数

$$W_k=\frac{2^{n+1}n!\sqrt{\pi}}{[H_{n+1}(x_k)]^2}$$

公式的截断误差是

$$E_n=\frac{n!\sqrt{\pi}}{2^n(2n)!}f^{(2n)}(\xi)\quad -\infty<\xi<\infty$$

计算埃尔米特多项式可以采用递推公式

$$H_{n+2}(x)=2xH_{n+1}(x)-2(n+1)H_n(x)$$

$$H_0(x)=1$$

$$H_1(x)=2x$$

关于这些积分公式的收敛性，已有定理证明了：

(a) 如果函数 $f(x)$ 当 x 充分大时满足不等式

$$|f(x)|\leqslant\frac{e^x}{x^{\alpha+1+\rho}}\quad 对某个\ \rho>0(\alpha>-1)$$

则广义拉盖尔积分公式收敛到积分

$$\lim_{n\to\infty}\sum_{k=1}^{n}W_kf(x_k)=\int_0^{\infty}e^{-x}x^{\alpha}f(x)dx$$

(拉盖尔积分公式是其 $\alpha=0$ 的特殊情形)。

(b) 如果函数 $f(x)$ 当 $|x|$ 充分大时，满足不等式

$$|f(x)|\leqslant\frac{e^{x^2}}{|x|^{1+\rho}}\quad 对某个\ \rho>0$$

则埃尔米特积分公式收敛到积分

$$\lim_{n\to\infty}\sum_{k=1}^{n}W_kf(x_k)=\int_0^{\infty}e^{-x^2}f(x)dx$$

四、柯西主值的计算

与求奇异积分柯西主值同样的分解方法，可得

$$P\int_{-\infty}^{\infty} f(x)dx = 2\int_{0}^{\infty}[f(x)+f(-x)]dx$$

§ 2.11 计算重积分的累次积分法

考虑二重积分

$$\iint_D f(x, y)dxdy \tag{2.11.1}$$

此处积分域 D 是平面上的一个有界域。常见的有如下二种区域:

(1) $a\leqslant x\leqslant b \quad c\leqslant y\leqslant d$

(2) $a\leqslant x\leqslant b \quad c(x)\leqslant y\leqslant d(x)$

此处 $c(x)$ 和 $d(x)$ 是 $[a, b]$ 上的连续函数,且有 $c(x)\leqslant d(x)$。

对于任何简单的平面有界域,总可以通过区域的适当分解,化为若干个这样区域的和。因此我们只讨论二重积分

$$I = \iint_{\substack{a\leqslant x\leqslant b\\ c(x)\leqslant y\leqslant d(x)}} f(x, y)dxdy \tag{2.11.2}$$

此处 $c(x)$, $d(x)$ 是 $[a, b]$ 上的连续函数,亦可为常数,且 $f(x, y)$ 在积分域上假定是连续的。

我们记

$$g(x) = \int_{c(x)}^{d(x)} f(x, y)dy \tag{2.11.3}$$

对于任何固定的 x 值,这是个单重定积分,因此函数 $g(x)$ 是有定义的,且根据 $c(x)$、$d(x)$ 及 $f(x, y)$ 的连续性,它也是连续的。于是积分

$$I = \int_a^b g(x)dx \tag{2.11.4}$$

存在。因此我们可以将二重积分(2.11.2)化为累次积分

$$I = \int_a^b \left[\int_{c(x)}^{d(x)} f(x, y)dy\right]dx \tag{2.11.5}$$

将一个多重积分化为累次积分后,任何前面介绍的求单重积分的数值积分法都可以用来求一多重积分。因为对于积分式(2.11.4),可以应用某个数值积分法来计算它的量值,即可写成

$$I \approx \sum_{i=1}^{n} W_i g(x_i) \tag{2.11.6}$$

用关系式(2.11.3)代入后,有

$$I \approx \sum_{i=1}^{n} W_i \int_{c(x_i)}^{d(x_i)} f(x_i, y)dy \tag{2.11.7}$$

对于式中的 n 个定积分

$$g(x_i) = \int_{c(x_i)}^{d(x_i)} f(x_i, y)dy \tag{2.11.8}$$

我们亦可以应用某个数值积分法算出它的量值

$$g(x_i) = \sum_{j=1}^{m_i} V_{ij} f(x_i, y_j) \tag{2.11.9}$$

以此代入(2.11.7)式,于是得积分近似式

$$I\approx\sum_{i=1}^{n}W_i\sum_{j=1}^{m_i}V_{ij}f(x_i, y_j) \qquad (2.11.10)$$

对于三重积分的累次积分法亦是类似的。但累次积分法仅适用于低重积分。对于高重积分，由于工作量太大，一般不宜使用。

§2.12 计算高重积分的数论网格法

假设要计算 S 重积分

$$\int_0^1\cdots\int_0^1 f(x_1, x_2, \cdots, x_S)dx_1, \cdots, dx_S \qquad (2.12.1)$$

此处积分域是 S 维单位超立方体。对于一般的积分域，可以通过域的分解和变量变换，化为在此标准域上的积分。对此有一系列的数论网格方法。本节仅介绍其中之一，即柯洛波夫(Коробов)方法[5]。

在介绍方法之前，先定义两个函数类：

(1) 设 $f(x_1, \cdots, x_S)$ 在 S 维超立方体

$$0\leqslant x_i\leqslant 1 \quad (i=1, 2, \cdots, S) \qquad (2.12.2)$$

上是连续的。且设 $f(x_1, \cdots, x_S)$ 对每个变量 $x_1, x_2, \cdots, x_S$ 都是以 1 为周期的周期函数。且令 $c(m_1, \cdots, m_S)$ 表示 f 的傅里叶系数，即

$$f(x_1, \cdots, x_S)=\sum_{m_1=-\infty}^{\infty}\cdots\sum_{m_S=-\infty}^{\infty}c(m_1, \cdots, m_S)\exp[2\pi i(m_1x_1+\cdots+m_Sx_S)] \qquad (2.12.3)$$

$$c(m_1, \cdots, m_S)=\int_0^1\cdots\int_0^1 f(x_1, \cdots, x_S)\exp[-2\pi i(m_1x_1+\cdots+m_Sx_S)]dx_1\cdots dx_S \qquad (2.12.4)$$

如果满足

$$|c(m_1, \cdots, m_S)|\leqslant\frac{c}{(\overline{m}_1\overline{m}_2\cdots\overline{m}_S)^{\alpha}} \qquad (2.12.5)$$

此处 $\overline{m}_i=\max(1, |m_i|)(i=1, \cdots, S)$，而 α 是大于 1 的实数，c 是与 $m_1, \cdots, m_S$ 无关的常数，则称 f 属于 $E_S^{\alpha}(c)$ 类的，记为 $f\in E_S^{\alpha}(c)$。

特别在 α 为整数时，具有连续导数 $\frac{\partial^{\alpha S}f}{\partial x_1^{\alpha}\cdots\partial x_S^{\alpha}}$ 的周期函数是属于 $E_S^{\alpha}(c)$ 类的。

(2) 如果 $f(x_1, \cdots, x_S)$ 在单位超立方体(2.12.2)上导数

$$\frac{\partial^n f}{\partial x_1^{n_1}\cdots\partial x_S^{n_S}} \quad (0\leqslant n\leqslant\alpha S, 0\leqslant n_i\leqslant\alpha) \quad \alpha\text{ 是}\geqslant 1\text{ 的整数} \qquad (2.12.6)$$

存在且连续，并且有界 c_1，则称 f 属于 $H_S^{\alpha}(c_1)$ 类，记为 $f\in H_S^{\alpha}(c_1)$。

对于积分(2.12.1)，考虑到积分公式

$$\int_0^1\cdots\int_0^1 f(x_1, \cdots, x_S)dx_1\cdots dx_S=\frac{1}{N}\sum_{k=1}^{N}f[\xi_1(k), \cdots, \xi_S(k)]-E \qquad (2.12.7)$$

此处点集合 $M=\{\xi_1(k), \cdots, \xi_S(k)\}$ 称为网格；E 是积分公式的截断误差。

如果把 S 维单位超立方体分成 $N=n^S$ 个相等的小超立方体，这就得到均匀网格。但对于 $E_S^{\alpha}(c)$ 类中的函数，其误差估计为

$$E=O\left(\frac{1}{N^{\alpha/S}}\right) \qquad (2.12.8)$$

由此看出，对于均匀网格，随着积分重数 S 的增加，其准确度迅速地下降。因此为了提高精

度，必须采用大量的网格点数，而使工作量增大到无法容忍的程度。

利用数论网格，可以克服这一缺点，可以证明(见[5])：

对于任何素数 $N>2$，存在整数 $a_i=a_i(N)$ $(i=1, 2, \cdots, S)$，使得对于任何 $f\in E_S^\alpha(c)$，求积公式

$$\int_0^1\cdots\int_0^1 f(x_1, \cdots, x_S)dx_1\cdots dx_S=\frac{1}{N}\sum_{k=1}^{N}f\left[\left\{\frac{ka_1}{N}\right\}, \cdots, \left\{\frac{ka_S}{N}\right\}\right]+E \tag{2.12.9}$$

有误差估式

$$E=O\left(\frac{\ln^{\alpha(S-1)}N}{N^\alpha}\right) \tag{2.12.10}$$

其中符号 $\{u\}$ 表示数 u 的小数部分，下同。

具有上述性质的整数 $a_1, \cdots, a_S$ 称之为按模 N 的最优系数。而点集合

$$M=\left\{\left\{\frac{ka_1}{N}\right\}, \left\{\frac{ka_2}{N}\right\}, \cdots, \left\{\frac{ka_S}{N}\right\}\right\}$$

称为超平行六面体网格。它是数论网格的一种。

应用数论网格法计算积分时，先要对取定的数 N，算出相应的最优系数。下面仅给出计算最优系数的具体方法，其理论根据就不予以讨论了。

最优系数的计算方法

方法 1　设 N 是大于 S 的素数，我们定义函数

$$H(z)=\frac{3^S}{N}\left[1+2\sum_{k=1}^{\frac{N-1}{2}}\left(1-2\left\{\frac{k}{N}\right\}\right)^2\left(1-2\left\{\frac{kz}{N}\right\}\right)^2\cdots\left(1-2\left\{\frac{kz^{S-1}}{N}\right\}\right)^2\right]$$

$$=\frac{3^S}{N}\left[1+2\sum_{k=1}^{\frac{N-1}{2}}\prod_{j=0}^{S-1}\left(1-2\left\{\frac{kz^j}{N}\right\}\right)^2\right] \tag{2.12.11}$$

如果在 $z=a$ 时，函数 $H(z)$ 在区间 $1\leqslant z\leqslant\frac{N-1}{2}$ 上达到极小，则整数 $1, a, \cdots, a^{S-1}$ 是按模 N 的最优系数。

用这种方法计算最优系数，需要作 $O(N^2)$ 次基本运算。当 N 比较大时，采用下述方法 2 可以减少工作量。

方法 2　令 $N=N_1N_2$，其中 N_1, N_2 是大于 S 的素数。则先按方法 1，对 $N=N_1$ 确定数 a，然后令

$$\widetilde{H}(z)=\frac{3^S}{N_2}\left[1+2\sum_{k=1}^{\frac{N_2-1}{2}}\prod_{j=0}^{S-1}\left(1-2\left\{k\,\frac{N_1z^j+N_2a^j}{N_2}\right\}\right)^2\right] \tag{2.12.12}$$

如果在 $z=b$ 时，使函数 $\widetilde{H}(z)$ 在区间 $1\leqslant z\leqslant\frac{N_2-1}{2}$ 上达到极小，则整数 $N_1+N_2, N_1b+N_2a, \cdots, N_1b^{S-1}+N_2a^{S-1}$ 是按模 N 的最优系数。用这个方法计算最优系数时，需要作 $O(N_1^2+N_1N_2^2)$ 次基本运算，当 $N_2\approx\sqrt{N_1}$ 时，其工作量为 $O(N^{1+\frac{1}{3}})$。

在本章附表中，我们已给出了 $3\leqslant S\leqslant 10$ 的一批最优系数。

上面对 $E_S^\alpha(c)$ 类的函数给出了使用数论网格的积分法。如果被积函数是非周期性的，但是属于 $H_S^\alpha(c_1)$ 类的，那末在使用该方法之前，需对 f 进行周期化处理。即由 $f(x_1, \cdots, x_S)$ 构造函数 $\varphi(x_1, \cdots, x_S)$，使有

(1) $\varphi(x_1, \cdots, x_S)$ 是对所有的变量 $x_1, \cdots, x_S$ 都是以 1 为周期的周期函数；

(2) $\int_0^1\cdots\int_0^1 f(x_1,\ \cdots,\ x_S)dx_1\cdots dx_S=\int_0^1\cdots\int_0^1\varphi(x_1,\cdots,x_S)dx_1\cdots dx_S$

下面提供几种函数周期化的方法：

第一种周期化方法：作

$$\left.\begin{aligned}&\varphi_1(x_1,\ \cdots,\ x_S)=\frac{1}{2}[f(x_1,\ x_2,\ \cdots,\ x_S)+f(1-x_1,\ x_2,\ \cdots,\ x_S)]\\&\varphi_2(x_1,\ \cdots,\ x_S)=\frac{1}{2}[\varphi_1(x_1,\ x_2,\ \cdots,\ x_S)+\varphi_1(x_1,\ 1-x_2,\ \cdots,\ x_S)]\\&\cdots\cdots\\&\varphi_S(x_1,\ \cdots,\ x_S)=\frac{1}{2}[\varphi_{S-1}(x_1,\ \cdots,\ x_{S-1},\ x_S)+\varphi_{S-1}(x_1,\ \cdots,\ x_{S-1},\ 1-x_S)]\end{aligned}\right\}\quad(2.12.13)$$

取 $\varphi(x_1,\ \cdots,\ x_S)=\varphi_S(x_1,\ \cdots,\ x_S)$。

第二种周期化方法：对所有的变量作变量变换

$$x=\tau(z)=3z^2-2z^3\qquad(2.12.14)$$

得

$$\varphi(x_1,\ x_2,\ \cdots,\ x_S)=f[\tau(x_1),\ \cdots,\ \tau(x_S)]\tau'(x_1)\cdots\tau'(x_S)\qquad(2.12.15)$$

可以证明，上两种周期化得到的 φ 满足条件(1)、(2)，且有 $\varphi\in E_S^2(c)$。故得积分公式

$$\int_0^1\cdots\int_0^1 f(x_1,\ \cdots,\ x_S)dx=\frac{1}{N}\sum_{k=1}^{N}\varphi\left(\left\{\frac{ka_1}{N}\right\},\ \cdots,\ \left\{\frac{ka_S}{N}\right\}\right)+E\qquad(2.12.16)$$

有误差估式

$$E=O\left(\frac{\ln^{2(S-1)}N}{N^2}\right)$$

但在第一种周期化方法中，计算一个 φ 值需执行 2^S 次 f 值的计算。因此在公式(2.12.16)中实际上需执行 2^SN 次 f 值的计算。这就显然降低了方法的有效性。

上两种周期化方法，仅对函数本身执行了周期化处理，使其满足条件(1)，而其导数一般说来不一定满足条件(1)，因此只能得到 $O\left(\frac{\ln^{2(S-1)}N}{N^2}\right)$ 级的误差。如要使 φ 的导数亦满足条件(1)，使有 $\varphi\in E_S^\alpha$，而获得与估式(2.12.10)同样的误差，则可以在第二种周期化方法中，将变换改为

$$\tau(x)=(2\alpha-1)C_{2(\alpha-1)}^{\alpha-1}\left(\frac{x^\alpha}{\alpha}-\frac{C_{\alpha-1}^1}{\alpha+1}x^{\alpha+1}+\cdots\pm\frac{C_{\alpha-1}^{\alpha-1}x^{2\alpha-1}}{2\alpha-1}\right)$$

此处，C_n^m 是组合符号，即

$$C_n^m=\frac{n!}{m!(n-m)!},$$

对于非周期的 H_S^α 类函数，也可以不作周期化处理，而直接应用公式(2.12.9)。但此时误差要大些，只有 $O\left(\frac{\ln^SN}{N}\right)$ 级的误差。

附表 2.1　高斯-勒让德求积公式的结点和系数

结点 $=\pm x_i$（勒让德多项式的零点）；

W_i——系数；n——结点数。

n	x_i	W_i
2	0.57735 02691 89626	1.00000 00000 00000
3	0.00000 00000 00000	0.88888 88888 88889
	0.77459 66692 41483	0.55555 55555 55556
4	0.33998 10435 84856	0.65214 51548 62546
	0.86113 63115 94053	0.34785 48451 37454
5	0.00000 00000 00000	0.56888 88888 88889
	0.53846 93101 05683	0.47862 86704 99366
	0.90617 98459 38664	0.23692 68850 56189
6	0.23861 91860 83197	0.46791 39345 72691
	0.66120 93864 66265	0.36076 15730 48139
	0.93246 95142 03152	0.17132 44923 79170
7	0.00000 00000 00000	0.41795 91836 73469
	0.40584 51513 77397	0.38183 00505 05119
	0.74153 11855 99394	0.27970 53914 89277
	0.94910 79123 42759	0.12948 49661 68870
8	0.18343 46424 95650	0.36268 37833 78362
	0.52553 24099 16329	0.31370 66458 77887
	0.79666 64774 13627	0.22238 10344 53374
	0.96028 98564 97536	0.10122 85362 90376
9	0.00000 00000 00000	0.33023 93550 01260
	0.32425 34234 03809	0.31234 70770 40003
	0.61337 14327 00590	0.26061 06964 02935
	0.83603 11073 26636	0.18064 81606 94857
	0.96816 02395 07626	0.08127 43883 61574
10	0.14887 43389 81631	0.29552 42247 14753
	0.43339 53941 29247	0.26926 67193 09996
	0.67940 95682 99024	0.21908 63625 15982
	0.86506 33666 88985	0.14945 13491 50581
	0.97390 65285 17172	0.06667 13443 08688
12	0.12523 34085 11469	0.24914 70458 13403
	0.36783 14989 98180	0.23349 25365 38355
	0.58731 79542 86617	0.20316 74267 23066
	0.76990 26741 94305	0.16007 83285 43346
	0.90411 72563 70475	0.10693 93259 95318
	0.98156 06342 46719	0.04717 53363 86512
16	0.09501 25098 37637 440185	0.18945 06104 55068 496285
	0.28160 35507 79258 913230	0.18260 34150 44923 588867
	0.45801 67776 57227 386342	0.16915 65193 95002 538189
	0.61787 62444 02643 748447	0.14959 59888 16576 732081
	0.75540 44083 55003 033895	0.12462 89712 55533 872052
	0.86563 12023 87831 743880	0.09515 85116 82492 784810
	0.94457 50230 73232 576078	0.06225 35239 38647 892863
	0.98940 09349 91649 932596	0.02715 24594 11754 094852
20	0.07652 65211 33497 333755	0.15275 33871 30725 850698
	0.22778 58511 41645 078080	0.14917 29864 72603 746788
	0.37370 60887 15419 560673	0.14209 61093 18382 051329
	0.51086 70019 50827 098004	0.13168 86384 49176 626898
	0.63605 36807 26515 025453	0.11819 45319 61518 417312
	0.74633 19064 60150 792614	0.10193 01198 17240 435037
	0.83911 69718 22218 823395	0.08327 67415 76704 748725
	0.91223 44282 51325 905868	0.06267 20483 34109 063570
	0.96397 19272 77913 791268	0.04060 14298 00386 941331
	0.99312 85991 85094 924786	0.01761 40071 39152 118312

附表2.2　高斯-拉盖尔积分公式的结点和系数

$$\int_0^\infty e^{-x}f(x)dx=\sum_{i=1}^{n}W_if(x_i)\qquad \int_0^\infty g(x)dx=\sum_{i=1}^{n}W_ie^{x_i}g(x_i)$$

结点$=x_i$（拉盖尔多项式的零点）；

W_i——系数；　n——公式结点数。

n	x_i			W_i			$W_ie^{x_i}$		
2	0.58578	64376	27	0.85355	33905	93	1.53332	60331	2
	3.41421	35623	73	0.14644	66094	07	4.45095	73350	5
3	0.41577	45567	83	0.71109	30099	29	1.07769	28592	7
	2.29428	03602	79	0.27851	77335	69	2.76214	29619	0
	6.28994	50829	37	0.10389	25650	16×10^{-1}	5.60109	46254	3
4	0.32254	76896	19	0.60315	41043	42	0.83273	91238	38
	1.74576	11011	58	0.35741	86924	38	2.04810	24384	5
	4.53662	02969	21	0.38887	90851	50×10^{-1}	3.63114	63058	2
	9.39507	09123	01	0.53929	47055	61×10^{-3}	6.48714	50844	1
5	0.26356	03197	18	0.52175	56105	83	0.67909	40422	08
	1.41340	30591	07	0.39866	68110	83	1.63848	78736	0
	3.59642	57710	41	0.75942	44968	17×10^{-1}	2.76944	32423	7
	7.08581	00058	59	0.36117	58679	92×10^{-2}	4.31565	69009	2
	12.64080	08442	76	0.23369	97238	58×10^{-4}	7.21918	63543	5
6	0.22284	66041	79	0.45896	46739	50	0.57353	55074	23
	1.18893	21016	73	0.41700	08307	72	1.36925	25907	1
	2.99273	63260	59	0.11337	33820	74	2.26068	45933	8
	5.77514	35691	05	0.10399	19745	31×10^{-1}	3.35052	45823	6
	9.83746	74183	83	0.26101	72028	15×10^{-3}	4.88682	68002	1
	15.98287	39806	02	0.89854	79064	30×10^{-6}	7.84901	59456	0
7	0.19304	36765	60	0.40931	89517	01	0.49647	75975	40
	1.02666	48953	39	0.42183	12778	62	1.17764	30608	6
	2.56787	67449	51	0.14712	63486	58	1.91824	97816	6
	4.90035	30845	26	0.20633	51446	87×10^{-1}	2.77184	86362	3
	8.18215	34445	63	0.10740	10143	28×10^{-2}	3.84124	91224	9
	12.73418	02917	98	0.15865	46434	86×10^{-4}	5.38067	82079	2
	19.39572	78622	63	0.31703	15479	00×10^{-7}	8.40543	24868	3
8	0.17027	96323	05	0.36918	85893	42	0.43772	34104	93
	0.90370	17767	99	0.41878	67808	14	1.03386	93476	7
	2.25108	66298	66	0.17579	49866	37	1.66970	97656	6
	4.26670	01702	88	0.33343	49226	12×10^{-1}	2.37692	47017	6
	7.04590	54023	93	0.27945	36235	23×10^{-2}	3.20854	09133	5
	10.75851	60101	81	0.90765	08773	36×10^{-4}	4.26857	55108	3
	15.74067	86412	78	0.84857	46716	27×10^{-6}	5.81808	33686	7
	22.86313	17368	89	0.10480	01174	87×10^{-8}	8.90622	62152	9

（续表）

n	x_i			W_i			$W_ie^{x_i}$		
9	0.15232	22277	32	0.33612	64217	98	0.39143	11243	16
	0.80722	00227	42	0.41121	39804	24	0.92180	50285	29
	2.00513	51556	19	0.19928	75253	71	1.48012	79099	4
	3.78347	39733	31	0.47460	56276	57×10^{-1}	2.08677	08075	5
	6.20495	67778	77	0.55996	26610	79×10^{-2}	2.77292	13897	1
	9.37298	52516	88	0.30524	97670	93×10^{-3}	3.59162	60680	9
	13.46623	69110	92	0.65921	23026	08×10^{-5}	4.64876	60021	4
	18.83359	77889	92	0.41107	69330	35×10^{-7}	6.21227	54197	5
	26.37407	18909	27	0.32908	74030	35×10^{-10}	9.36321	82377	1
10	0.13779	34705	40	0.30844	11157	65	0.35400	97386	07
	0.72945	45495	03	0.40111	99291	55	0.83190	23010	44
	1.80834	29017	40	0.21806	82876	12	1.33028	85617	5
	3.40143	36978	55	0.62087	45609	87×10^{-1}	1.86306	39031	1
	5.55249	61400	64	0.95015	16975	18×10^{-2}	2.45025	55580	8
	8.33015	27467	64	0.75300	83885	88×10^{-3}	3.12276	41551	4
	11.84378	58379	00	0.28259	23349	60×10^{-4}	3.93415	26955	6
	16.27925	78313	78	0.42493	13984	96×10^{-6}	4.99241	48721	9
	21.99658	58119	81	0.18395	64823	98×10^{-8}	6.57220	24851	3
	29.92069	70122	74	0.99118	27219	61×10^{-12}	9.78469	58403	7
12	0.11572	21173	58	0.26473	13710	55	0.29720	96360	44
	0.61175	74845	15	0.37775	92758	73	0.69646	29804	31
	1.51261	02697	76	0.24408	20113	20	1.10778	13946	2
	2.83375	13377	44	0.90449	22221	17×10^{-1}	1.53846	42390	4
	4.59922	76394	18	0.20102	38115	46×10^{-1}	1.99832	76062	7
	6.84452	54531	15	0.26639	73541	87×10^{-2}	2.50074	57691	0
	9.62131	68424	57	0.20323	15926	63×10^{-3}	3.06532	15182	8
	13.00605	49933	06	0.83650	55856	82×10^{-5}	3.72328	91107	8
	17.11685	51874	62	0.16684	93876	54×10^{-6}	4.52981	40299	8
	22.15109	03793	97	0.13423	91030	52×10^{-8}	5.59725	84618	4
	28.48796	72509	84	0.30616	01635	04×10^{-11}	7.21299	54609	3
	37.09912	10444	67	0.81480	77467	43×10^{-15}	10.54383	74619	
15	0.09330	78120	17	0.21823	48859	40	0.23957	81703	11
	0.49269	17403	02	0.34221	01779	23	0.56010	08427	93
	1.21559	54120	71	0.26302	75779	42	0.88700	82629	19
	2.26994	95262	04	0.12642	58181	06	1.22366	44021	5
	3.66762	27217	51	0.40206	86492	10×10^{-1}	1.57444	87216	3
	5.42533	66274	14	0.85638	77803	61×10^{-2}	1.94475	19765	3
	7.56591	62266	13	0.12124	36147	21×10^{-2}	2.34150	20566	4
	10.12022	85680	19	0.11167	43923	44×10^{-3}	2.77404	19268	3
	13.13028	24821	76	0.64599	26762	02×10^{-5}	3.25564	33464	0
	16.65440	77083	30	0.22263	16907	10×10^{-6}	3.80631	17142	3
	20.77647	88994	49	0.42274	30384	98×10^{-8}	4.45847	77538	4
	25.62389	42267	29	0.39218	97267	04×10^{-10}	5.27001	77844	3
	31.40751	91697	54	0.14565	15264	07×10^{-12}	6.35956	34697	3
	38.53068	33064	86	0.14830	27051	11×10^{-15}	8.03178	76321	2
	48.02608	55726	86	0.16005	94906	21×10^{-19}	11.52777	21009	

附表2.3 高斯-埃尔米特求积公式的结点和系数

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx = \sum_{i=1}^{n} W_i f(x_i) \qquad \int_{-\infty}^{\infty} g(x)dx = \sum_{i=1}^{n} W_i e^{x_i^2} g(x_i)$$

结点$=\pm x_i$(埃尔米特多项式的零点);

W_i——系数; n——公式的结点数。

n	x_i	W_i	$W_i e^{x_i^2}$
2	0.70710 67811 86548	0.88622 69254 528	1.46114 11826 611
3	0.00000 00000 00000	0.11816 35900 604×10¹	1.18163 59006 037
	1.22474 48713 91589	0.29540 89751 509	1.32393 11752 136
4	0.52464 76232 75290	0.80491 40900 055	1.05996 44828 950
	1.65068 01238 85785	0.81312 83544 725×10⁻¹	1.24022 58176 958
5	0.00000 00000 00000	0.94530 87204 829	0.94530 87204 829
	0.95857 24646 13819	0.39361 93231 522	0.98658 09967 514
	2.02018 28704 56086	0.19953 24205 905×10⁻¹	1.18148 86255 360
6	0.43607 74119 27617	0.72462 95952 244	0.87640 13344 362
	1.33584 90740 13697	0.15706 73203 229	0.93558 05576 312
	2.35060 49736 74492	0.45300 09905 509×10⁻²	1.13690 83326 745
7	0.00000 00000 00000	0.81026 46175 568	0.81026 46175 568
	0.81628 78828 58965	0.42560 72526 101	0.82868 73032 836
	1.67355 16287 67471	0.54515 58281 913×10⁻¹	0.89718 46002 252
	2.65196 13568 35233	0.97178 12450 995×10⁻³	1.10133 07296 103
8	0.38118 69902 07322	0.66114 70125 582	0.76454 41286 517
	1.15719 37124 46780	0.20780 23258 149	0.79289 00483 864
	1.98165 67566 95843	0.17077 98300 741×10⁻¹	0.86675 26065 634
	2.93063 74202 57244	0.19960 40722 114×10⁻³	1.07193 01442 480
9	0.00000 00000 00000	0.72023 52156 061	0.72023 52156 061
	0.72355 10187 52838	0.43265 15590 026	0.73030 24527 451
	1.46855 32892 16668	0.88474 52739 438×10⁻¹	0.76460 81250 946
	2.26658 05845 31843	0.49436 24275 537×10⁻²	0.84175 27014 787
	3.19099 32017 81528	0.39606 97726 326×10⁻⁴	1.04700 35809 767
10	0.34290 13272 23705	0.61086 26337 353	0.68708 18539 513
	1.03661 08297 89514	0.24013 86110 823	0.70329 63231 049
	1.75668 36492 99882	0.33874 39445 548×10⁻¹	0.74144 19319 436
	2.53273 16742 32790	0.13436 45746 781×10⁻²	0.82066 61264 048
	3.43615 91188 37738	0.76404 32855 233×10⁻⁵	1.02545 16913 657
12	0.31424 03762 54359	0.57013 52362 625	0.62930 78743 695
	0.94778 83912 40164	0.26049 23102 642	0.63962 12320 203
	1.59768 26351 52605	0.51607 98561 588×10⁻¹	0.66266 27732 669
	2.27950 70805 01060	0.39053 90584 629×10⁻²	0.70522 03661 122
	3.02063 70251 20890	0.85736 87043 588×10⁻⁴	0.78664 39394 633
	3.88972 48978 69782	0.26585 51684 356×10⁻⁶	0.98969 90470 923

(续表)

n	x_i	W_i	$W_i e^{x_i^2}$
16	0.27348 10461 3815	0.50792 94790 166	0.54737 52050 378
	0.82295 14491 4466	0.28064 74585 285	0.55244 19573 675
	1.38025 85391 9888	0.83810 04139 899×10^{-1}	0.56321 78290 882
	1.95178 79909 1625	0.12880 31153 551×10^{-1}	0.58124 72754 009
	2.54620 21578 4748	0.93328 40086 242×10^{-3}	0.60973 69582 560
	3.17699 91619 7996	0.27118 60092 538×10^{-4}	0.65575 56728 761
	3.86944 79048 6012	0.23209 80844 865×10^{-6}	0.73824 56222 777
	4.68873 89393 0582	0.26548 07474 011×10^{-9}	0.93687 44928 841
20	0.24534 07083 009	0.46224 36696 006	0.49092 15006 667
	0.73747 37285 454	0.28667 55053 628	0.49384 33852 721
	1.23407 62153 953	0.10901 72060 200	0.49992 08713 363
	1.73853 77121 166	0.24810 52088 746×10^{-1}	0.50967 90271 175
	2.25497 40020 893	0.32437 73342 238×10^{-2}	0.52408 03509 486
	2.78880 60584 281	0.22833 86360 163×10^{-3}	0.54485 17423 644
	3.34785 45673 832	0.78025 56478 532×10^{-5}	0.57526 24428 525
	3.94476 40401 156	0.10860 69370 769×10^{-6}	0.62227 86961 914
	4.60368 24495 507	0.43993 40992 273×10^{-9}	0.70433 29611 769
	5.38748 08900 112	0.22293 93645 534×10^{-12}	0.89859 19614 532

注：以上三个表引自参考资料[12]。在该处，对高斯-勒让德积分公式还给出了最高到 $n=96$ 的结点和权系数。

附表 2.4 数论网格法的最优系数[6]

S——积分重数；

$N=p$（素数）——公式的结点数；

$a_i(i=1,2,\cdots)$——最优系数。对所有的情况，$a_1\equiv1$；

$H(b)-1$——误差标志数。对于 $f\in E_S^\alpha(c)$，积分公式有误差估式 $|E|\leqslant c\left(\frac{\pi^2}{6}\right)^{\frac{\alpha S}{2}}[H(b)-1]^{\frac{\alpha}{2}}$。

	$S=3$			$S=4$			
$N=p$	$H(b)-1$	a_2	a_3	$H(b)-1$	a_2	a_3	a_4
101	0.0703	40	85				
199	0.0214	30	104				
307	0.0114	75	99	0.0906	42	229	101
523	0.00454	78	331	0.0412	178	304	243
701	0.00319	215	660	0.0281	82	415	382
1069	0.00142	136	323	0.0150	71	765	865
1543	0.00075	355	1042	0.00837	128	954	215
2129	0.00044	359	1141	0.00500	766	1281	1906
3001	0.00025	276	1151	0.00303	174	266	1269
4001	0.00015	722	1154	0.00200	113	766	2537
5003	0.000105	1476	2274	0.001480	792	1889	191
6007	0.000070	592	2058	0.001009	1351	5080	3086
8191	0.000044	739	5515	0.000622	2488	5939	7859
10007	0.000033	544	5733	0.000486	1206	3421	2842
20039	0.000016	5704	12319	0.000276	19668	17407	14600
28117	0.000008	19449	5600	0.000108	17549	1900	24455
39029	0.000005	10607	26871	0.000077	30699	34367	60605
57091	0.000002	48188	21101	0.000056	52590	48787	38790
82001	0.000001	21252	67997	0.000031	57270	58903	17672
100063	0.000001	28036	22431	0.000019	92513	24700	95582

（续表）

	$S=5$					$S=6$					
$N=p$	$H(b)-1$	a_2	a_3	a_4	a_5	$H(b)-1$	a_2	a_3	a_4	a_5	a_6
1069	0.0962	63	762	970	177						
1543	0.0580	58	278	694	134						
2129	0.0383	618	833	1705	1964	0.186	41	1681	793	578	279
3001	0.0237	408	1409	1681	1620	0.123	233	271	122	1417	51
4001	0.0154	1534	568	3095	2544	0.086	1751	1235	1945	844	1475
5003	0.0114	840	177	3593	1311	0.063	2037	1882	1336	4803	2846
6007	0.0085	509	780	558	1693	0.050	312	1232	5943	4060	5250
8191	0.0055	1386	4302	7715	3735	0.034	1632	1349	6380	1399	6070
10007	0.0042	198	9183	6967	8507	0.027	2240	4093	1908	931	3984
15019	0.0032	10641	2640	6710	784	0.0197	8743	8358	6559	2795	772
20039	0.0022	11327	11251	12076	18677	0.0123	5557	150	11951	2461	9179
33139	0.0011	32133	17866	21281	32247	0.0069	18236	1831	19143	5522	22910
51097	0.0006	44672	45346	7044	14242	0.0031	9931	7551	29682	44446	17340
71053	0.0003	33755	65170	12470	6878	0.0026	18010	3155	50203	6605	13328
100063	0.0002	90036	77477	27253	6222	0.0015	43307	15440	39114	43534	39955

	$S=7$						
$N=p$	$H(b)-1$	a_2	a_3	a_4	a_5	a_6	a_7
15019	0.0835	12439	2983	8607	7041	7210	6741
18101	0.0730	17487	14946	44	9186	7308	1936
24041	0.0463	1833	18190	21444	23858	1135	12929
33139	0.0339	7642	9246	5584	23035	32241	30396
46213	0.0210	37900	17534	41873	32280	15251	26909
57091	0.0168	35571	45299	51436	34679	1472	8065
71053	0.0131	31874	36082	13810	6605	68784	9848
100063	0.0085	39040	62047	89839	6347	30892	64404

	$S=8$							
$N=p$	$H(b)-1$	a_2	a_3	a_4	a_5	a_6	a_7	a_8
24041	0.1999	17471	21749	5411	12326	3144	21024	6252
33139	0.1345	3520	29553	3239	1464	16735	19197	3019
46213	0.0900	5347	30775	35645	11403	16894	32016	16600
57091	0.0688	17411	46802	9779	16807	35302	1416	47755
71053	0.0557	60759	26413	24409	48215	51048	19876	29096
100063	0.0359	4344	58492	29291	60031	10486	22519	60985

	$S=9$								
$N=p$	$H(b)-1$	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
33193	0.4915	68	4624	16181	6721	26221	26661	23442	3384
46213	0.3262	8871	40115	20065	30352	15654	42782	17966	33962
57091	0.2664	20176	12146	23124	2172	33475	5070	42339	36122
71053	0.2021	26454	13119	27174	17795	22805	43500	45665	49857
100063	0.1136	70893	53211	12386	27873	56528	16417	17628	14997
159053	0.0846	60128	101694	23300	43576	57659	42111	85501	93062

（续表）

	S=10									
$N=p$	$H(b)-1$	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
85633	0.614	37667	35345	43864	54821	74078	30354	57935	51906	56297
103661	0.499	45681	57831	80987	9718	51556	55377	37354	4353	27595
115069	0.431	65470	650	95039	77293	98366	70366	74605	55507	49201
130703	0.377	64709	53373	17385	5244	29008	52889	66949	51906	110363
145087	0.333	55464	120722	105045	102309	58342	5327	59596	60510	119243
155093	0.316	90485	20662	110048	102308	148396	125399	124635	10480	44198

附录 积 分 程 序

一、梯形积分法(给定步长)程序
TRAP(a, b, h, I, FUNC)

使用说明

过程TRAP是应用梯形积分公式计算定积分$I=\int_a^b f(x)dx$。a, b是积分限; h是积分步长; I是积分计算值; 过程FUNC是计算被积函数值$f(x)$的, 需有过程说明: 过程FUNC(x, f); 值x; 简变f; 始……终;

```
过程 TRAP(A, B, H, I, FUNC);
  值 A, B, H;
  简变 I;
  过程 FUNC;
始
  简变 N, F, FB, X;
    §ENTI((B-A)/H)⇒N; A⇒X; FUNC(X, F); F/2⇒I;
    对于 J=1 到 N 步长 1 执行
    始 X+H⇒X; FUNC(X, F); I+F⇒I
    终;
    H*(I-F/2)⇒I; FUNC(B, FB); (B-X)*(F+FB)/2+I⇒I;
终;
```

二、辛浦生积分法(给定步长)程序
SIMP(a, b, h, I, FUNC)

使用说明

过程SIMP是应用辛浦生求积公式计算定积分$I=\int_a^b f(x)dx$。a, b是积分限; h是积分步长; I是积分计算值; 过程FUNC是计算被积函数值$f(x)$的, 需有过程说明: 过程FUNC(x, f); 值x; 简变f; 始……终;

```
过程 SIMP(A, B, H, I, FUNC);
  值 A, B, H;
  简变 I;
  过程 FUNC;
```

```
始
  简变 X, F, N, FB;
    §ENTI((B-A)/(2*H))⇒N; A⇒X; FUNC(A, I);
    对于 J=1 到 N 步长 1 执行
    始 X+H⇒X; FUNC(X, F); I+4*F⇒I; X+H⇒X; FUNC(X, F); I+2*F⇒I;
    终;
     H*(I-F)/3⇒I; FUNC((X+B)/2, FB); F+4*FB⇒F; FUNC(B, FB);
     (B-X)*(F+FB)/6+I⇒I;
终;
```

三、辛浦生积分法(自动选步长)程序
ASMP(a, b, eps, I, FUNC)

使用说明

过程 ASMP 是用辛浦生积分法计算定积分 $I=\int_a^b f(x)dx$。过程按精度要求自动选定积分步长。

输入参数:

a, b: 积分限;

eps: 允许误差;

FUNC: 计算被积函数值 $f(x)$ 的过程,需有过程说明:过程 FUNC(x, f); 值 x; 简变 f; 始……终;

输出参数:

I: 积分计算值。

```
过程 ASMP(A, B, EPS, I, FUNC);
  值 A, B, EPS;
  简变 I;
  过程 FUNC;
始
  简变 N, H, S, S1, F, I1, E;
    2⇒N; (B-A)/2⇒H; FUNC(A, S); FUNC(B, F); S+F⇒S;
  ITRT: 0⇒S1;
    对于 K=1 到 N 步长 2 执行
    始 FUNC(A+K*H, F); S1+F⇒S1;
    终;
    (S+4*S1)*H/3⇒I; S+2*S1⇒S;
    若 N=2
    则否始 I-I1⇒E;
              若 §ABS(I)<1 则否 E/I⇒E;
```

```
        若§ABS(E)<EPS 则转 OUT 否;
      终;
  I⇒I1; 2*N⇒N; H/2⇒H;
  转 ITRT;
 OUT:
终;
```

四、逐次分半加速积分法程序

ROMB(a, b, eps, mink, maxk, I, FUNC, FAIL)

使用说明

过程ROMB是应用逐次分半加速积分法计算定积分$I=\int_a^b f(x)dx$。

输入参数:

a, b: 积分限;

eps: 允许误差;

min$k$: 2^{mink}是最低限度的区间分段数,是为防止假收敛之用;

max$k$: 2^{maxk}是最高限度的区间分段数,是为控制工作量之用;

FUNC: 计算被积函数值$f(x)$的过程,过程需有说明:过程 FUNC(x, f);值x;简变f;始……终;

FAIL: 非正常出口,当在区间分段数$\leqslant 2^{maxk}$的范围内不能达到指定的精度要求时,将转向FAIL[1]。

输出参数:

I: 积分计算值。

```
过程 ROMB(A, B, EPS, min K, max K, I, FUNC, FAIL);
 值 A, B, EPS, min K, max K;
 简变 I;
 过程 FUNC;
 开关 FAIL;
始
 简变 K, H, S, X, F, DT, M;
 场 T[0:max K];
   0⇒K; B-A⇒H; FUNC(A, F); F⇒S; FUNC(B, F); (S+F)/2⇒S; H*S⇒T[0];
 ITRT: K+1⇒K; H/2⇒H;
   对于 J=1 到 2↑K-1 步长 2 执行
   始 FUNC(A+J*H, F); S+F⇒S 终;
   H*S⇒T[K]; T[K]-T[K-1]⇒DT;
   若 §ABS(T[K])<1
```

```
    则否 DT/T[K]⇒DT;
    若 §ABS(DT)<EPS
    则若 min K≤K 则始 T[K]⇒T[0]; 转 OUT 终否
    否;
    1⇒M;
    对于 J=K-1 到 0 步长 -1 执行
    始 T[J+1]-T[J]⇒DT; 0.25*M⇒M; T[J]+DT/(1-M)⇒T[J]
    终;
    若 §ABS(T[0])<1
    则否 DT/T[0]⇒DT;
    若 §ABS(DT)<EPS
    则若 K<min K 则转 ITRT 否转 OUT
    否若 K<maxK 则转 ITRT 否转 FAIL[1];
  OUT: T[0]⇒I;
终;
```

五、用切氏级数展开的积分法(计算定积分)程序 CLEN(a, b, eps, maxN, I, FUNC, FAIL)

使用说明

过程 CLEN 是用切氏级数展开的积分法计算定积分 $I=\int_a^b f(x)\,dx$。

输入参数:

- a, b: 积分限;
- *eps*: 允许误差;
- max$N$: 允许的最大结点数,是为控制工作量之用;
- FUNC: 计算被积函数值 $f(x)$ 的过程,需有过程说明;
- FAIL: 非正常出口,当在结点数 ≤max$N$ 的范围内达不到精度要求时,将转向 FAIL[1];

输出参数:

- I: 积分计算值。

```
过程 CLEN(A, B, EPS, maxN, I, FUNC, FAIL);
  值 A, B, EPS, maxN;
  简变 I;
  过程 FUNC;
  开关 FAIL;
始
  简变 PAI, T, X, F, C1, C0, N, M;
```

```
场 C[0:maxN],FF[1:maxN];
过程 CA;
始
  简变 A0, AI, S, PAIN;
  PAI/(N+N)⇒PAIN;
  对于 I=0 到 0.5*N 步长 1 执行
  始 0.5*C[I]⇒C[I]; C[I]⇒C[N-I];
  终;

  0⇒A0;
  对于 J=1 到 N 步长 1 执行
  始 §COS((J+J-1)*PAIN)⇒T; C1*T+C0⇒X;
     FUNC(X, F); F⇒FF[J]; A0+F⇒A0;
  终;

  A0/N⇒A0; C[0]+A0⇒C[0]; C[N]-A0⇒C[N];
  对于 I=1 到 0.5*N-1 步长 1 执行
  始 0⇒AI; 2*I*PAIN⇒S;
    对于 J=1 到 N 步长 1 执行
      AI+FF[J]*§COS(S*(J+J-1))⇒AI;  AI/N⇒AI;  C[I]+AI⇒C[I];
      C[N-I]-AI⇒C[N-I];
  终
终;

0.5*(B-A)⇒C1; 0.5*(B+A)⇒C0;
3.1415926⇒PAI;
FUNC(A, F); F⇒FF[1];
FUNC(0.5*(A+B), F); F⇒FF[2];
FUNC(B, F); F⇒FF[3];
0.5*(FF[1]+FF[3])⇒F;
F+FF[2]⇒C[0];
F-FF[2]⇒C[1];
2⇒N;
ITRT: CA; 0.5*C[0]⇒I;
对于 J=1 到 N 步长 1 执行
  I-C[J]/(4*J*J-1)⇒I; I*(B-A)⇒I; (C[N-2]-C[N-1])/(4*N-6)⇒M;
(C[N-1]-C[N])/(4*N-2)⇒T;
若 §ABS(M)<§ABS(T) 则 T⇒M 否;
C[N]/(4*N+2)⇒T;
若 §ABS(M)<§ABS(T) 则 T⇒M 否;
```

```
  若 §ABS(I)<1 则否 M/I⇒M;
  若 §ABS(M)<EPS 则转 OUT 否;
  N+N⇒N;
  若 MAXN<N 则转 FAIL[1]否转 ITRT;
  OUT:
终;
```

六、用切氏级数展开的积分法(计算不定积分)程序 ITGL(eps, maxN, b, N, FUNC, FAIL)

使用说明

过程 ITGL 是用切氏级数展开的积分法给出不定积分 $I(x)=\int_{-1}^{x} f(x)dx^{**}(-1<x\leqslant 1)$ 的近似表达式: $I(x)\sim\frac{b_0}{2}+\sum_{i=1}^{N} b_i T_i(x)$, 此处$T_i(x)=\cos(i\cos^{-1}x)$是切比雪夫多项式。过程给出表达式中的系数 $b_0, b_1, b_2, \cdots, b_N$。

输入参数:

eps: 允许误差;

max$N$: 允许的最大 N 数(近似式中的项数),用以控制工作量的;

FUNC: 计算被积函数值 $f(x)$的过程,需有过程说明: 过程 FUNC(x, f); 值 x; 简变 f; 始……终;

FAIL: 非正常出口,当在 $N\leqslant$max$N$ 的范围内达不到精度要求时,将转向 FAIL[1];

输出参数:

b: 依次赋值近似表达式中的系数 $b_0, b_1, \cdots, b_N$;

N: 项数。

** 对于积分 $I(t)=\int_a^t f(t)\,dt, a<t\leqslant b$,可施以变量变换 $t=\frac{(b-a)x+(b+a)}{2}$, 变成积分 $\frac{b-a}{2}\int_{-1}^{x} f\left(\frac{(b-a)x+(b+a)}{2}\right)dx, -1<x\leqslant 1$, 而得表达式: $I(t)\sim\frac{b-a}{2}\left[\frac{b_0}{2}+\sum_{i=1}^{N} b_i T_i(x)\right]$, 此处 $x=\frac{2t-(b+a)}{b-a}$。

```
过程 ITGL(EPS, MAXN, B, N, FUNC, FAIL);
  值 EPS, MAXN;
  场 B;
  简变 N;
  过程 FUNC;
  开关 FAIL;
始
  简变 PAI, X, F, I, M;
  场 A[0:MAXN], FF[1:MAXN/2];
```

```
过程 CA;
始
  简变 A0, A1, AI, N2, T, PAIN;
  0.5*N⇒N2; PAI/N⇒PAIN;
  对于 I=0 到 N2 步长 1 执行
  始 0.5*A[I]⇒A[I]; A[I]⇒A[N-I];
  终;
  0⇒A0; 0⇒A1;
  对于 J=1 到 N2 步长 1 执行
  始 §COS((J+J-1)*PAIN)⇒X; FUNC(X, F); F⇒FF[J]; A0+F⇒A0;
    A1+F*X⇒A1;
  终;
  A0/N2⇒A0; A1/N2⇒A1; A[0]+A0⇒A[0]; A[N]-A0⇒A[N]; A[1]+A1⇒A[1];
  A[N-1]-A1⇒A[N-1];
  对于 I=2 到 N2-1 步长 1 执行
  始 0⇒AI; I*PAIN⇒T;
    对于 J=1 到 N2 步长 1 执行
      AI+FF[J]*§COS(T*(J+J-1))⇒AI; AI/N2⇒AI; a[I]+AI⇒A[I];
      A[N-I]-AI⇒A[N-I];
  终
终;
过程 CB;
始
  对于 I=1 到 N-1 步长 1 执行
    (A[I-1]-A[I+1])/(I+I)⇒B[I]; A[N-1]/(N+N)⇒B[N];
    A[N]/(N+N+2)⇒B[N+1];
  对于 I=0 到 N 步长 2 执行
    B[0]-B[I]+B[I+1]⇒B[0]; B[0]+B[0]⇒B[0];
终;
3.1415926⇒PAI;
FUNC(1, F); F⇒FF[1];
FUNC(0, F); f⇒FF[2];
FUNC(-1, F); F⇒FF[3];
0.5*(FF[1]+FF[3])⇒F;
F+FF[2]⇒A[0];
0.5*(FF[1]-FF[3])⇒A[1];
F-FF[2]⇒A[2];
2⇒N;
ITRT: CB;
```

```
  0.5*B[0]⇒I:
  对于 J=1 到 N+1 步长 1 执行
    I+B[J]⇒I; B[N-1]⇒M;
  若 §ABS(M)<§ABS(B[N])  则 B[N]⇒M 否;
  若 §ABS(M)<§ABS(B[N+1])  则 B[N+1]⇒M 否;
  若 §ABS(I)<1 则否 M/I⇒M;
  若 §ABS(M)<EPS
  则始 N+1⇒N; 转 OUT 终否;
  N+N⇒N;
  若 MAXN<N+1 则转 FAIL[1]否;
  CA;
  转 ITRT.
  OUT:
终;
```

七、计算切比雪夫级数值程序

CHEB(N, x, I, b)

使用说明

过程 CHEB 是计算 N 次切比雪夫级数 $\frac{b_0}{2}+\sum_{i=1}^{N} b_i T_i(x)$ 在 x 处的量值 I，此处 $T_i(x)=\cos(i\cos^{-1}x)$ 是切比雪夫多项式。

输入参数:

N: 切比雪夫级数的次数;

x: 自变量值;

b: 以切氏级数的系数 $b_0, b_1, \cdots, b_N$ 组成的 $N+1$ 元向量;

输出参数:

I: 切氏级数在 x 处的量值。

```
过程 CHEB(N, X, I, B);
  值 N, X;
  简变 I;
  场 B;
始
  简变 C0, C1, C2, TWOX; X+X⇒TWOX; B[N]⇒C0; 0⇒C1;
      对于 J=N-1 到 1 步长 -1 执行
      始 C1⇒C2; C0⇒C1; TWOX*C1-C2+B[J]⇒C0;
      终;
      X*C0-C1+0.5*B[0]⇒I;
终;
```

八、在离散点上给出函数的积分程序

NITG(a, b, n, xF, I)

使用说明

过程 NITG 是应用平均抛物插值法计算由离散点上给出的函数

$$f(x)=\begin{pmatrix} x_1, & x_2, & \cdots, & x_n \\ f_1, & f_2, & \cdots, & f_n \end{pmatrix}$$

的积分 $I=\int_a^b f(x)\,dx$。

输入参数:

a, b: 积分限(要求 $a<x_n$, $b>x_1$);

n: 给出函数值的离散点个数;

XF: $f(x)$的数表。$2n$ 元的一维场,依次赋值: x_1, f_1, x_2, f_2, $\cdots$, x_n, f_n。

输出参数:

I: 积分值。

```
过程 NITG(A, B, N, XF, I);
 值 A, B, N;
 场 XF;
 简变 I;
始
 简变 H, DLT, LMD, L, R, I0, K, K0, KL, C, IK;
 过程 LR(S);
   值 S;
 始
   S+S-1⇒K; XF[K+2]-XF[K]⇒H;
   若 2≤S
   则始 H/(XF[K]-XF[K-2])⇒DLT;
        DLT/(1+DLT)*(DLT*XF[K-1]-(1+DLT)*XF[K+1]+XF[K+3])⇒L;
     终
   否;
   若 S≤N-2
   则始 H/(XF[K+4]-XF[K+2])⇒LMD;
        LMD/(1+LMD)*(XF[K+1]-(1+LMD)*XF[K+3]+LMD*XF[K+5])⇒R;
     终
   否 L⇒R;
   若 S=1 则 R⇒L 否;
   H*(0.5*(XF[K+1]+XF[K+3])-(L+R)/12)⇒IK;
```

```
终;
过程 ENDI(X, S);
  值 X, S;
始
  简变 W;
  LR(S); (X-XF[K])/H⇒W; (X-XF[K])/6*(((L+R)*W-3*(0.5*(L+R)
    +XF[K+1]-XF[K+3]))*W+6*XF[K+1])⇒I0;
终;
若 A=B 则始 0⇒I; 转 OUT 终 否;
若 A<B
则 1⇒C
否始 A⇒C; B⇒A; C⇒B; -1⇒C 终;
对于 J=1 到 N 步长 1 执行
    始 J⇒K0;
      若 A⩽XF[J+J-1] 则转 N1 否;
    终;
转 N4;
N1: 若 K0=1 则 2⇒K0 否;
对于 J=1 到 N 步长 1 执行
    始 J-1⇒KL;
      若 B<XF[J+J-1] 则转 N2 否;
    终;
N2: 若 KL=0 则转 N4 否;
若 KL<K0 则转 N4 否;
ENDI(A, K0-1); IK-I0⇒I;
对于 K=K0 到 KL-1 步长 1 执行
    始 LR(K); I+IK⇒I;
    终;
ENDI(B, KL); C*(I+I0)⇒I;
转 OUT;
N4: 0⇒I;
OUT;
终;
```

参 考 资 料

[1] Davis, P. J. and Rabinowitz, P.: Numerical Integration. Blaisdell Publishing Company, 1967.

[2] Romberg, W.: Vereinfachte Numerische Integration. Norske Vid. Selsk. Forh. Trondheim, 28 (1955), pp. 30-36.

[3] Bauer, F. L. Rutishauser, H. and stiefel, E.: New Aspects in Numerical Integration. Proceedings of Symposia of Applied Mathematics, Vol. 15, Amer. Math. Soc., 1963, pp. 199-218.

[4] Clenshaw, C.W. and Curtis, A. R.: A Method for Numerical Integration on an Automatie Computer. Num. Math, Vol. 2 (1960), pp. 197-205.

[5] 华罗庚与王元,《数值积分及其应用》,科学出版社,1963.

[6] А. И. Салтыков: Журнал Вычислительной Математики и Математической Физики Том3, № 1, 1963, pp. 181-186.

[7] A. Ralston and H. S. Wilf, "Mathematical Methods for Digital Computers", Volume II, John Wiley & Sons, New york, 1967.

[8] A. H. Stroud, "Error Estimates for Romberg Quadrature", J. SIAM Series B: Numer. Analysis, Vol. 2, No. 3, 1965, pp. 480-488.

[9] W. Fraser and W. Wilson,"Remarks on the Clenshaw-Curtis Quardrature Scheme", SIAM Review, Vol. 8, No. 3, 1966, pp. 322-327.

[10] J. P. Imhof,"On the Method for Numerical Integration of Clenshaw and Curtis", Numer. Math., Vol. 5, 1963, pp. 138-141.

[11] H. O'Hara and F. J. Smith, "Error Estimation in the Clenshaw-Curtis Quadrature formula", Computer J. Vol. 11, No. 2, 1968, pp. 213-219.

[12] M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables". National Bureau of Standards Appl. Math. Series. 55 (1966).

[13] 计算方法,北京大学、吉林大学、南京大学计算数学教研室编,1961.

第三章 谐波分析

自然界中种种复杂的振动现象，是由许多不同频率、不同振幅的简谐振动迭加而来；也就是说，一个复杂的波形可以分解为一系列的谐波。例如，光波为不同强度、不同波长的单色光，形成光谱。声音振动可以分解为不同音调、不同音强的“声谱”。无线电天线从空间接收不同来源的不同频率、不同振幅的电磁波，在天线回路中形成极其复杂的迭加波形。人们运用分光镜、耳鼓以及收音机的调谐线路，能够分别对这些系统进行析谱，这就是用物理的手段进行谐波分析。

针对这类现象，也发展了一套有效的数学工具，即傅里叶变换方法。它包括解析的方法（如傅氏级数、傅氏积分）和代数的方法（如有限傅氏级数，即离散傅氏变换），这些就是谐波分析的数学手段。

§3.1 傅氏级数

一个简谐振动可以表为

$$f(t)=A\sin(2\pi\nu t+\alpha)=A\sin(2\pi t/T+\alpha)=a\cos 2\pi t/T+b\sin 2\pi t/T$$

式中 $\nu=\frac{1}{T}$ 为频率；T 为周期；A 为振幅；α 为初始位相角。基频 $\nu=\frac{1}{T}$ 的各阶倍频 $k\nu=\frac{k}{T}$ $(k=0, 1, 2, \cdots)$ 的简谐振动的迭加，显然仍具有周期 T（例如图 3.1）。因此一般地可以设

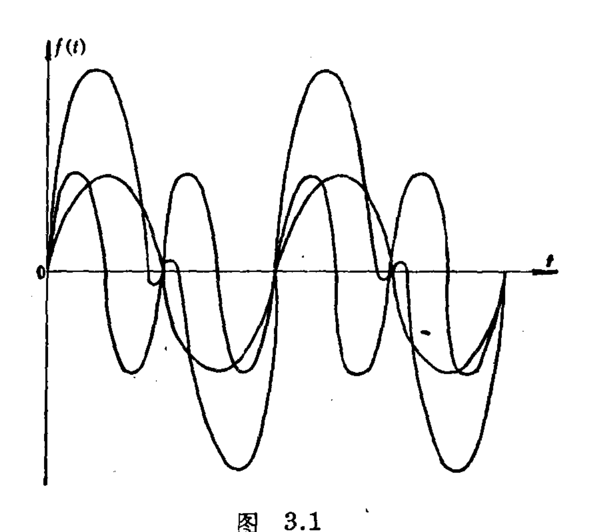


图 3.1

想任何以 T 为周期的振动，即周期函数

$$f(t)\equiv f(t+T),\quad -\infty<t<\infty \tag{3.1.1}$$

可以表为这些简谐振动的无穷迭加，即表为傅氏级数

$$f(t)=\frac{a_0}{2}+\sum_{k=1}^{\infty}a_k\cos 2\pi kt/T+\sum_{k=1}^{\infty}b_k\sin 2\pi kt/T,\quad -\infty<t<\infty \tag{3.1.2}$$

也可以表为复数形式

$$f(t)=\sum_{k=-\infty}^{\infty}c_k e^{2\pi ikt/T},\quad -\infty<t<\infty \tag{3.1.3}$$

不同形式下傅氏系数的关系是

$$c_0=\frac{a_0}{2},\quad c_k=\frac{1}{2}(a_k-ib_k),\quad c_{-k}=\frac{1}{2}(a_k+ib_k),\quad k=0,\ 1,\ \cdots$$

$$a_k=c_k+c_{-k},\quad b_k=i(c_k-c_{-k}) \tag{3.1.4}$$

为了讨论的方便,以下主要采取复数形式。

为了定出傅氏系数,将(3.1.3)两端各乘以 $e^{-2\pi ik't/T}$ 并对一个周期积分

$$\int_0^T f(t)\,e^{-2\pi ik't/T}dt=\sum_{k=-\infty}^{\infty}c_k\int_0^T e^{2\pi i(k-k')t/T}dt$$

根据三角函数在基本周期上的正交性

$$\int_0^T e^{2\pi ijt/T}dt=\begin{cases}T, & 当\ j=0\\ 0, & 当\ j\neq 0\end{cases} \tag{3.1.5}$$

可知,前式右端一切 $k\neq k'$ 的项均为 0, 而只剩下一项 $c_{k'}T$, 因此就得傅氏系数的表达式

$$c_k=\frac{1}{T}\int_0^T f(t)e^{-2\pi ikt/T}dt,\qquad k=0,\ \pm1,\ \cdots \tag{3.1.6}$$

鉴于 $f(t)$ 及 $e^{-2\pi ikt/T}$ 都具有周期 T, 因此在任何长度为 T 的区间$[a,\ a+T]$上求积的结果是一样的,所以有

$$c_k=\frac{1}{T}\int_0^T f(t)e^{-2\pi ikt/T}dt=\frac{1}{T}\int_{-T/2}^{T/2}f(t)e^{-2\pi ikt/T}dt$$

$$=\frac{1}{T}\int_a^{a+T}f(t)e^{-2\pi ikt/T}dt,\qquad k=0,\ \pm1,\ \cdots \tag{3.1.7}$$

当$f(t)$是偶函数,即$f(t)\equiv f(-t)$时,恒有 $c_k=c_{-k}$, 因此可以表为余弦级数

$$f(t)=\frac{a_0}{2}+\sum_{k=1}^{\infty}a_k\cos 2\pi kt/T$$

当$f(t)$为奇函数,即$f(t)\equiv -f(-t)$时,恒有 $c_k=-c_{-k}$, 因此可以表为正弦级数

$$f(t)=\sum_{k=1}^{\infty}b_k\sin 2\pi kt/T$$

通常人们说的在某个区间$[a,\ a+T]$上将某个函数f展为傅氏级数是指以系数

$$c_k=\frac{1}{T}\int_a^{a+T}f(t)e^{-2\pi ikt/T}dt \tag{3.1.8}$$

形成傅氏级数

$$\sum_{k=-\infty}^{\infty}c_k e^{2\pi ikt/T} \tag{3.1.9}$$

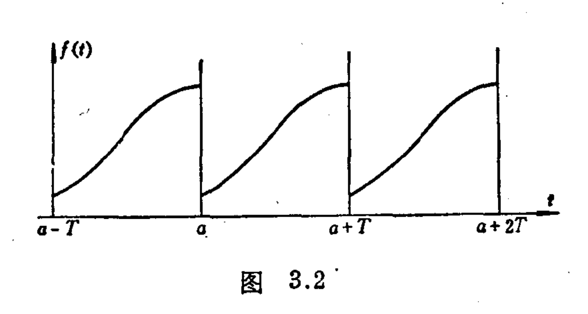


图 3.2

注意,这个级数在原区间$[a,\ a+T]$上确实等于原来那个函数$f(t)$。但是由于这个级数的周期性,它在全轴 $-\infty<t<\infty$ 上所表达的,则应该是由那段定义在 $[a,\ a+T]$ 上的函数以周期T向两端无穷延拓的结果,如图3.2。因此,即使原来的函数 $f(t)$ 在区间 $[a,\ a+T]$ 之外也是有意义的,但这一部分在形成级数(3.1.8),(3.1.9)时被截去了。

例1　在$\left[-\frac{1}{2}, \frac{1}{2}\right]$上将$f(t)=t$展为傅氏级数，得

$$t=\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{\pi k}\sin 2\pi kt, \quad |t|\leqslant\frac{1}{2}$$

它在$\left[-\frac{1}{2}, \frac{1}{2}\right]$上等于$t$，但在全轴上则等于图3.3所示的锯齿状间断函数。

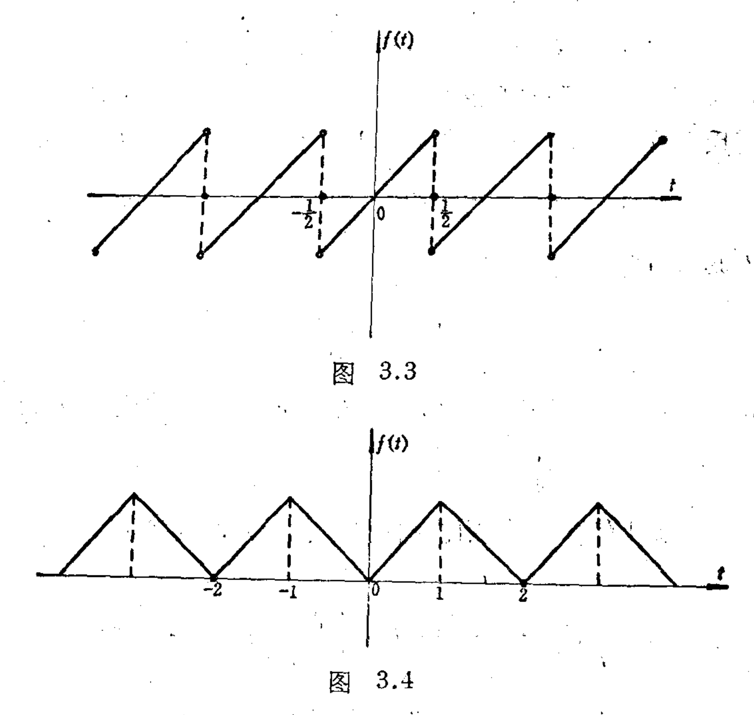


图　3.3

图　3.4

例2　在$[-1, 1]$上将$f(t)=|t|$展为傅氏级数，得

$$|t|=\frac{1}{2}-\frac{4}{\pi^2}\sum_{k=1}^{\infty}\frac{\cos(2k-1)\pi t}{(2k-1)^2}, \quad |t|\leqslant 1$$

它在$[-1, 1]$上等于$|t|$，而在全轴上则等于图3.4所示的锯齿状连续函数。

周期函数的光滑性与傅氏系数的行为有密切关系。例1，2所举的周期函数的光滑性依序递增，相应傅氏系数c_k随$|k|$增大的下降率愈甚，即级数的收敛速度递增。在一般情况下这一结论也成立，即可以证明，当周期函数作为整体——即考虑全轴$-\infty<t<\infty$——具有连续的m阶导数，则有

$$c_k\approx O\left(\frac{1}{|k|^{m+2}}\right), \quad 当\ |k|\to\infty \tag{3.1.10}$$

这个渐近估计有助于在近似计算中选择够多的项数。应该强调，这里的光滑性是从全轴来看的。例如，函数$f(t)\equiv t$在基本区间上是解析的，即无限光滑的，但由于将它以周期1无穷延拓出去后在半整数点$t=n+\frac{1}{2}$，$n=0, \pm1, \pm2, \cdots$有间断（图3.3），因此相应的收敛速度是很慢的。

当周期函数$f(t)$在某点$t=a$有间断（由于周期性，它必然在$t=a+mT$，$m=0, \pm1, \cdots$各点也有间断），可以证明，在这些点上傅氏级数的"值"等于$\frac{1}{2}(f(a+0)+f(a-0))$。因此，即使$f(t)$在某个区间$(a, a+T)$是连续的，但在两端点取不等的值，即$f(a)\neq f(a+T)$，

于是，相应傅氏级数在两端点的值同为 $\frac{1}{2}(f(a)+f(a+T))$，从例1及图3.3也表达了这一事实，相应的傅氏级数在 $t=\pm\frac{1}{2}$ 取值为 $0=\frac{1}{2}-\frac{1}{2}$。

§3.2 傅氏积分

把周期性振动分解为一系列简谐振动的迭加的主要特点是：它的“频谱”是由一个基频 $\Delta s=\frac{1}{T}$(T 为原周期)的整倍数 $s_k=\frac{k}{T}=k\Delta s$ $(k=0, 1, \cdots)$ 组成，即为等距的离散“谱线”。在非周期性振动，例如太阳光谱，它的谱线之间不一定有倍数关系，而且除了线状谱即离散谱以外，还有连续谱。因此，一般的非周期函数 $f(t)$ 在 $-\infty<t<\infty$ 上应展为具有更为一般形式的谐波迭加和，即“积分和”：

$$f(t)=\int_{-\infty}^{\infty}F(s)e^{2\pi ist}ds, \quad -\infty<t<\infty \tag{3.2.1}$$

这里 $F(s)$ 起着傅氏系数 c_k 的作用，但已经不是仅依赖于整数变量 k 而是作为连续变量 s 的函数，$-\infty<s<\infty$，即在频谱中一切实数频率都可能出现。

事实上，定义于 $-\infty<t<\infty$ 上的一般函数 $f(t)$ 可以视为周期函数在周期 $T\to\infty$ 时的极限。首先在区间 $-\frac{T}{2}\leqslant t\leqslant\frac{T}{2}$ 上展为傅氏级数

$$f(t)=\sum_{k=-\infty}^{\infty}c_k e^{2\pi ikt/T}, \quad -\frac{T}{2}\leqslant t\leqslant\frac{T}{2}$$

$$c_k=\frac{1}{T}\int_{-T/2}^{T/2}f(t)e^{-2\pi ikt/T}dt, \quad k=0, \pm1, \cdots$$

命 $\Delta s=\frac{1}{T}$，$\frac{k}{T}=k\Delta s=s_k$，$F(s_k)=Tc_k$，于是，当 $T\to\infty$ 时

$$f(t)=\sum_{k=-\infty}^{\infty}F(k\Delta s)e^{2\pi ik\Delta st}\Delta s=\sum_{k=-\infty}^{\infty}F(s_k)e^{2\pi is_kt}\Delta s\approx\int_{-\infty}^{\infty}F(s)e^{2\pi ist}ds$$

这里的函数 $F(s)$ 来自 $F(s_k)$，即当 $T\to\infty$ 时

$$F(s_k)=\int_{-T/2}^{T/2}f(t)e^{-2\pi is_kt}dt\approx F(s)=\int_{-\infty}^{\infty}f(t)e^{-2\pi ist}dt$$

这样就导致了一对互逆的傅氏积分

$$f(t)=\int_{-\infty}^{\infty}F(s)e^{2\pi ist}ds, \quad -\infty<t<\infty \tag{3.2.2}$$

$$F(s)=\int_{-\infty}^{\infty}f(t)e^{-2\pi ist}dt, \quad -\infty<s<\infty \tag{3.2.3}$$

$F(s)$ 称为 $f(t)$ 的傅氏变换或谱函数，$f(t)$ 称为 $F(s)$ 的逆傅氏变换或原函数。这个互逆关系将记为 $f(t)\sim F(s)$。如果变量 t 在物理上表示时间，则变量 s 代表频率，傅氏变换就建立了所谓“时间域”与“频率域”之间的互换关系。如果 t 表示空间坐标，则变量 s 表示距离 2π 内含有的波数，即 $\frac{2\pi}{s}$ 表示波长。

3.2.1 傅氏变换的基本性质

根据定义式(3.2.2)，(3.2.3)，傅氏正变换与逆变换的实质是相同的，只是积分号下指

数函数 $e^{\pm 2\pi ist}$ 的幂次反号。因此只须讨论正变换的性质，而逆变换的性质可以类似地导出。

首先，傅氏变换是线性的。事实上，从定义式(3.2.3)直接可以算出：

$$f(t)\sim F(s),\quad g(t)\sim G(s)\Rightarrow \alpha f(t)+\beta g(t)\sim \alpha F(s)+\beta G(s)$$

α, β 为任意常数，也可以是复数。

此外，设 $f(t)\sim F(s)$ 则有

$$F(t)\sim f(-s) \tag{3.2.4}$$

$$f(-t)\sim F(-s) \tag{3.2.5}$$

$$f^*(t)\sim F^*(-s),\quad {}^*\text{表示取复共轭} \tag{3.2.6}$$

$$f(t-a)\sim e^{-2\pi ias}F(s),\quad a\text{ 为任意实数} \tag{3.2.7}$$

$$f(at)\sim \frac{1}{|a|}F\left(\frac{s}{a}\right),\quad a\text{ 为任意实数}\neq 0 \tag{3.2.8}$$

$$f'(t)\sim 2\pi isF(s),\quad f^{(n)}(t)\sim(2\pi is)^nF(s) \tag{3.2.9}$$

事实上，根据互逆关系(3.2.2)，(3.2.3)可以算出(3.2.4)：

$$F(t)\sim\int_{-\infty}^{\infty}F(t)e^{-2\pi ist}dt=\int_{-\infty}^{\infty}F(t)e^{2\pi i(-s)t}dt=f(-s)$$

这是傅氏变换互逆性的另一种表达形式，即对函数 $f(t)$ 先作一次正变换，再作一次逆变换则还原为 $f(t)$，但是对 $f(t)$ 连作两次正变换则“反向还原”为 $f(-t)$。至于(3.2.5)到(3.2.8)则直接从定义式(3.2.3)并利用积分变量代换可以导出：

$$\begin{aligned}f(-t)\sim\int_{-\infty}^{\infty}f(-t)e^{-2\pi ist}dt&=\int_{-\infty}^{\infty}f(\tau)e^{2\pi is\tau}d\tau\\&=\int_{-\infty}^{\infty}f(\tau)e^{-2\pi i(-s)\tau}d\tau=F(-s)\end{aligned}$$

$$\begin{aligned}f^*(t)\sim\int_{-\infty}^{\infty}f^*(t)e^{-2\pi ist}dt&=\int_{-\infty}^{\infty}f^*(t)\left(e^{2\pi ist}\right)^*dt\\&=\int_{-\infty}^{\infty}\left[f(t)e^{-2\pi i(-s)t}\right]^*dt=F^*(-s)\end{aligned}$$

$$\begin{aligned}f(t-a)\sim\int_{-\infty}^{\infty}f(t-a)e^{-2\pi ist}dt&=\int_{-\infty}^{\infty}f(\tau)e^{-2\pi is(\tau+a)}d\tau\\&=\int_{-\infty}^{\infty}e^{-2\pi ias}f(\tau)e^{-2\pi is\tau}d\tau=e^{-2\pi ias}F(s)\end{aligned}$$

$$\begin{aligned}f(at)\sim\int_{-\infty}^{\infty}f(at)e^{-2\pi ist}dt&=\int_{-\infty}^{\infty}f(\tau)e^{-2\pi is\tau/a}\frac{d\tau}{|a|}\\&=\frac{1}{|a|}\int_{-\infty}^{\infty}f(\tau)e^{-2\pi i(s/a)\tau}d\tau=\frac{1}{|a|}F\left(\frac{s}{a}\right)\end{aligned}$$

对于(3.2.9)，则运用分部积分：

$$\begin{aligned}f'(t)\sim\int_{-\infty}^{\infty}f'(t)e^{-2\pi ist}dt&=\left[f(t)e^{-2\pi ist}\right]_{t=-\infty}^{t=+\infty}-\int_{-\infty}^{\infty}f(t)\frac{d}{dt}e^{-2\pi ist}dt\\&=0-\int_{-\infty}^{\infty}f(t)(-2\pi is)e^{-2\pi ist}dt=2\pi isF(s)\end{aligned}$$

将此式连用 n 遍就得到(3.2.9)的第二式。

当原函数 $f(t)$ 变为 $f(-t)$ 时，相应的图线变为反向；公式(3.2.5)表示谱函数 F 也变为反向。

根据傅氏变换的定义式(3.2.2 和 3.2.3)可以得到

$$f(0)=\int_{-\infty}^{\infty} F(s)ds \tag{3.2.10}$$

$$F(0)=\int_{-\infty}^{\infty} f(t)dt \tag{3.2.11}$$

设原函数 f 为偶函数，即具有对称性

$$f(t)\equiv f(-t)$$

根据式(3.2.5)可知谱函数 F 也有对称性

$$F(s)\equiv F(-s)$$

也是偶函数。此外，由于

$$F(s)=\int_{-\infty}^{\infty} f(t)e^{-2\pi ist}dt=\int_{-\infty}^{\infty} f(t)(\cos 2\pi st-i\sin 2\pi st)dt$$

$$=\int_{-\infty}^{\infty} f(t)\cos 2\pi st\,dt-i\int_{-\infty}^{\infty} f(t)\sin 2\pi st\,dt$$

利用 f 的对称性可知

$$\int_{-\infty}^{\infty} f(t)\cos 2\pi st\,dt=2\int_{0}^{\infty} f(t)\cos 2\pi st\,dt$$

$$\int_{-\infty}^{\infty} f(t)\sin 2\pi st\,dt=0$$

由此可知偶函数 f 的谱函数 F 可以表为“余弦变换”的形式

$$F(s)=2\int_{0}^{\infty} f(t)\cos 2\pi st\,dt \tag{3.2.12}$$

类似地，当 f 为奇函数，即具有反对称性

$$f(t)\equiv -f(-t)$$

则其谱函数 F 也是奇函数

$$F(s)\equiv -F(-s)$$

并且可以表为“正弦变换”的形式

$$F(s)=-2i\int_{0}^{\infty} f(t)\sin 2\pi st\,dt \tag{3.2.13}$$

实践上多数情况原函数 f 是实函数，即

$$f(t)=f^*(t)$$

它的谱函数一般不再是实函数，但是，根据(3.2.6)有

$$F(s)=F^*(-s)$$

即具有共轭对称性。据此可以导出：当 f 为实偶函数时，F 也是实偶函数；当 f 为实奇函数时，F 则是纯虚奇函数。这从余弦及正弦变换式(3.2.12)，(3.2.13)也可以看出。

函数 $f(t)$ 的“伸缩”表为 $f(at)$。例如，当 $a>1$ 时，$f(at)$ 相当于把 $f(t)$ 的图线沿轴压缩 a 倍；当 $0<a<1$ 时相当于沿轴拉伸 $\frac{1}{a}$ 倍。公式(3.2.8)表示原函数作尺度伸缩时，谱函数有相反相成的伸缩关系。由此蕴涵的意义将在 3.3.4 节中再说。

函数 $f(t)$ 的“平移”表为 $f(t-a)$。当 $a>0$ 时，$f(t-a)$ 相当于把 $f(t)$ 的图线向右平移一个距离 a，当 $a<0$ 时相当于向左平移。公式(3.2.7)表示原函数 $f(t)$ 作平移时，谱函数 $F(s)$ 被乘以相应的虚指数因子 $e^{-2\pi ias}$。在这个基础上，傅氏变换可以应用解差分方程。

公式(3.2.9)表示原函数 $f(t)$ 的微分运算相当于谱函数 $F(s)$ 被乘以 $2\pi is$，在这个基础

上傅氏变换可以应用于解微分方程。

上述公式(3.2.4 到 3.2.9)对于谱函数的实际计算也很有用。事实上,一旦算出了某个原函数的谱函数后,则根据这些公式就能得到由此派生的许多函数的谱函数,而无待于直接的计算。

我们将傅氏变换连同逆变换的基本性质列为表 3.1, 其中关于卷积以及内积的性质将在以后再说,为了参考方便故一并列入。

表 3.1 傅氏变换性质简表

	原函数	谱函数	说明
	$f(t)$ $g(t)$	$F(s)$ $G(s)$	$-\infty<t,\ s<+\infty$
1	$F(t)$	$f(-s)$	互逆性
2	$\alpha f(t)+\beta g(t)$	$\alpha F(s)+\beta G(s)$	线性
3	$f(-t)$	$F(-s)$	偶(奇)性～偶(奇)性
4	$f^*(t)$	$F^*(-s)$	实性～共轭对称性
5	$f(at)$ $\frac{1}{\lvert a\rvert}f\left(\frac{t}{a}\right)$	$\frac{1}{\lvert a\rvert}F\left(\frac{s}{a}\right)$ $F(as)$	尺度伸缩原理
6	$f(t-a)$ $e^{2\pi iat}f(t)$	$e^{-2\pi ias}F(s)$ $F(s-a)$	平移 a～乘因子 $e^{-2\pi ias}$
7	$f^{(n)}(t)$ $(-2\pi it)^n f(t)$	$(2\pi is)^n F(s)$ $F^{(n)}(s)$	微分～乘因子 $2\pi is$
8	$f(t)*g(t)=\int_{-\infty}^{\infty}f(\tau)g(t-\tau)d\tau$ $f(t)\cdot g(t)$	$F(s)G(s)$ $F(s)*G(s)=\int_{-\infty}^{\infty}F(\sigma)G(s-\sigma)d\sigma$	卷积～乘积
9	$\int_{-\infty}^{\infty}f^*(t)g(t)dt=\int_{-\infty}^{\infty}F^*(s)G(s)ds$ $\int_{-\infty}^{\infty}\lvert f(t)\rvert^2dt=\int_{-\infty}^{\infty}\lvert F(s)\rvert^2ds$ $\sum_{n=-\infty}^{\infty}f(nT)=\frac{1}{T}\sum_{n=-\infty}^{\infty}F\left(\frac{n}{T}\right)$		内积不变性 巴色瓦公式 泊松公式

顺便指出,傅氏变换的定义及记号在资料中很不统一。许多资料中傅氏变换规定为

$$F(s)=\int_{-\infty}^{\infty}f(t)e^{-ist}dt$$

这时逆变换则为

$$f(t)=\int_{-\infty}^{\infty}F(s)e^{ist}ds$$

即指数虚幂中省去常因子 2π, 这时在公式(3.2.7), (3.2.9)中也应作相应省略,缺点是正逆公式比较不对称。当然这些区别都是非本质的,但会引起一些混淆,在实践中要注意到这一点。

傅氏变换可以推广到多元函数。例如,对于二元函数有二维傅氏变换

$$F(u, v)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x, y)e^{-2\pi i(ux+vy)}dx\,dy$$

$$f(x, y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}F(u, v)e^{2\pi i(ux+vy)}du\,dv$$

这也是线性变换，而公式(3.2.4 到 3.2.9)则推广为：当 $f(x, y)\sim F(u, v)$ 则有

$$F(x, y)\sim f(-u, -v)$$

$$f(-x, -y)\sim F(-u, -v)$$

$$f^*(x, y)\sim F^*(-u, -v)$$

$$f(x-a, y-b)\sim e^{-2\pi i(au+bv)}F(u, v)$$

$$f(ax, by)\sim \frac{1}{|ab|}F\left(\frac{u}{a}, \frac{v}{b}\right)$$

$$\frac{\partial^{p+q}}{\partial x^p\partial y^q}f(x, y)\sim(2\pi i)^{p+q}u^pv^qF(u, v)$$

3.2.2 一些初等函数的傅氏变换

为了掌握运用傅氏方法，需要熟悉一些最简单又是最基本的"初等函数"及其谱函数。

矩形函数

$$\operatorname{rect} t=\begin{cases}1, & |t|\leqslant\dfrac{1}{2}\\ 0, & 它处\end{cases} \tag{3.2.14}$$

其中心在原点 $t=0$，长宽各为 1，因此面积为 1

$$\int_{-\infty}^{\infty}\operatorname{rect} t\,dt=\int_{-1/2}^{1/2}dt=1 \tag{3.2.15}$$

的矩形分布，有间断点 $t=\pm1/2$，如图 3.5。

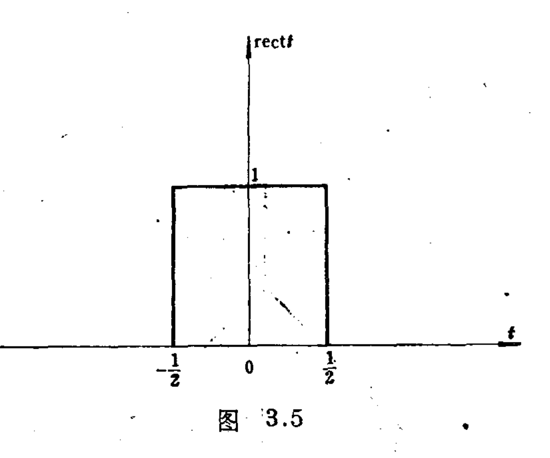


图 3.5

不难直接算出其谱函数

$$\int_{-\infty}^{\infty}\operatorname{rect} t\,e^{-2\pi ist}dt=\int_{-1/2}^{1/2}e^{-2\pi ist}dt$$

$$=\frac{\sin\pi s}{\pi s}\equiv\operatorname{sinc} s, -\infty<s<\infty \tag{3.2.16}$$

这就导致谐波分析中另一个重要的函数，即所谓 sinc 函数。

sinc 函数

$$\operatorname{sinc} t=\frac{\sin\pi t}{\pi t}, \quad -\infty<t<\infty \tag{3.2.17}$$

这是中峰在原点，左右各有逐渐衰减的正负"边瓣"的振动函数，见图 3.6。

$$\operatorname{sinc} 0=1, \operatorname{sinc} n=0, n=\pm1, \pm2, \cdots$$

$$\operatorname{sinc} t=0\left(\frac{1}{|t|}\right), \quad |t|\sim\infty \tag{3.2.18}$$

根据(3.2.10 到 3.2.11)有

$$\int_{-\infty}^{\infty}\operatorname{sinc} t\,dt=1 \tag{3.2.19}$$

即"面积"为 1。总结起来有

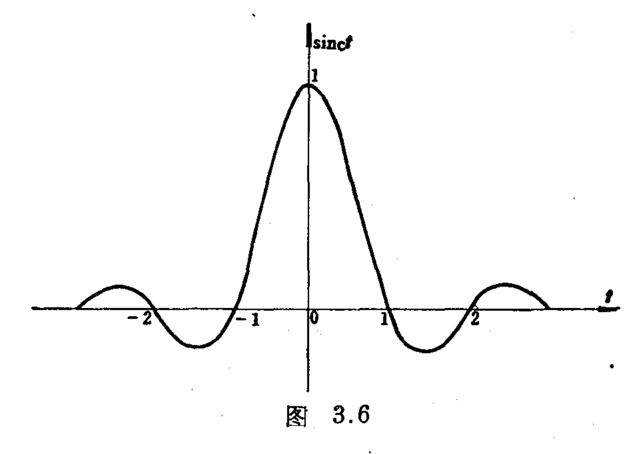


图 3.6

$$\text{rect}\,t \sim \text{sinc}\,s, \quad \text{sinc}\,t \sim \text{rect}\,s \tag{3.2.20}$$

三角形函数

$$\Lambda(t) = \begin{cases} 1-|t|, & |t| \leqslant 1 \\ 0, & \text{它处} \end{cases} \tag{3.2.21}$$

这是中心在原点，面积为1的等腰三角形分布，是连续函数，但一阶导数有间断($t=0, \pm 1$)，如图3.7。

$$\int_{-\infty}^{\infty} \Lambda(t)\,dt = \int_{-1}^{1} (1-|t|)\,dt = 1$$

也不难直接算出它的谱函数

$$\int_{-\infty}^{\infty} \Lambda(t) e^{-2\pi i s t}\,dt = \int_{-1}^{1} (1-|t|) e^{-2\pi i s t}\,dt$$

$$= \left(\frac{\sin \pi s}{\pi s}\right)^2 = \sin c^2 s,$$

如图3.8。

即
$$\Lambda(t) \sim \text{sinc}^2 s, \quad \text{sinc}^2 t \sim \Lambda(s) \tag{3.2.22}$$

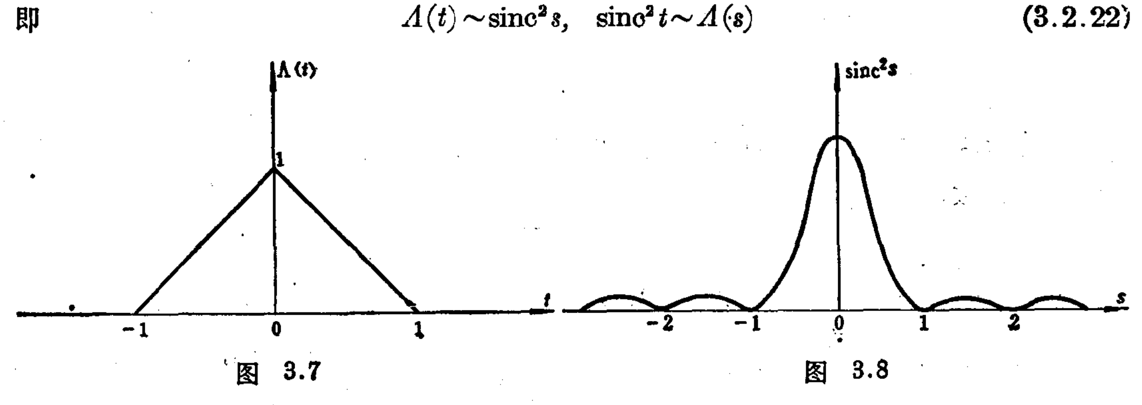


图 3.7 图 3.8

脉冲函数

取函数列

$$f_n(t) = n\,\text{rect}\,nt = \begin{cases} n, & \text{当 } |t| \leqslant \dfrac{1}{2n} \\ 0, & \text{它处} \end{cases} \quad n=1, 2, \cdots \tag{3.2.23}$$

它表示一系列愈来愈高，愈来愈窄，而面积保持为1，即

$$\int_{-\infty}^{\infty} f_n(t)\,dt=1$$

的矩形分布，如图3.9(a)。当$n\to\infty$时，它们的极限函数，记为$\delta(t)$，显然具有下列“奇异”的性质

$$\delta(t)=\begin{cases}0, & t\neq 0\\ \infty, & t=0\end{cases} \tag{3.2.24}$$

$$\int_{-\infty}^{\infty}\delta(t)\,dt=1 \tag{3.2.25}$$

函数$\delta(t)$叫做脉冲函数或δ函数。它表示位于原点的单位集中力、点电荷、点源以及单位强度的无穷窄脉冲等等理想化的物理量。上述极限过程可以写成

$$\lim_{n\to\infty} n\,\mathrm{rect}\,nt=\delta(t), \quad -\infty<t<\infty \tag{3.2.26}$$

如图3.9(b)，函数$\delta(t)$象征性地表为一个位于原点的箭头，其方向及长度表示其积分值$+1$。

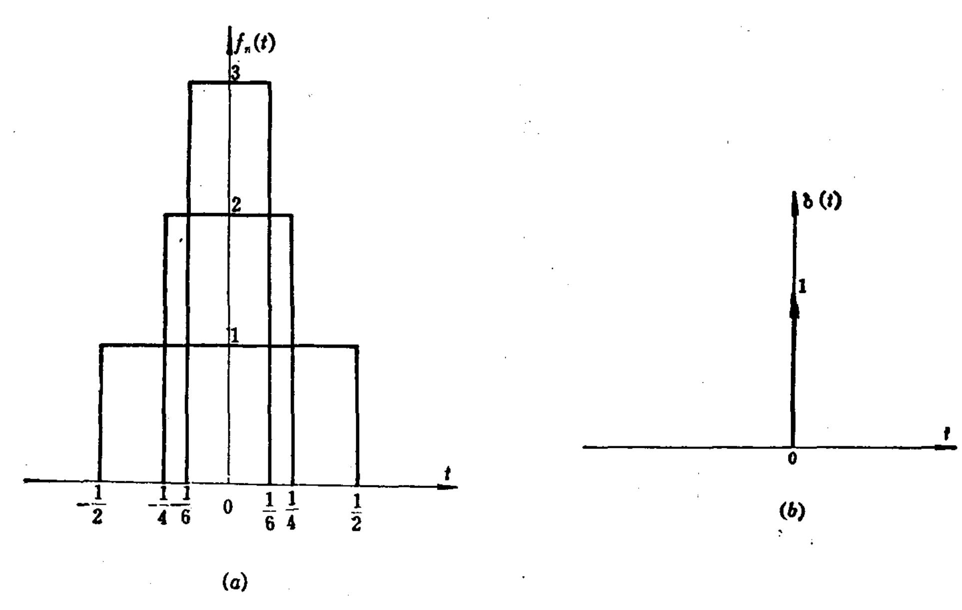


图 3.9

脉冲函数的筛取作用：

任取连续函数$\varphi(t)$，由于

$$\int_{-\infty}^{\infty}(n\,\mathrm{rect}\,nt)\varphi(t)\,dt=\int_{-1/2n}^{1/2n} n\varphi(t)\,dt\approx\varphi(0)\int_{-1/2n}^{1/2n} n\,dt=\varphi(0), \quad n\to\infty$$

因此有

$$\int_{-\infty}^{\infty}\delta(t)\varphi(t)\,dt=\varphi(0) \tag{3.2.27}$$

对一切连续函数$\varphi(t)$成立。这就是δ函数的筛取作用，也是δ函数的最基本的，即特征的性质。δ函数尽管有高度的“奇异性”，但是由于式(3.2.27)，它的运算却是极其简单的，甚至比普通的函数远为简单。

将函数$\delta(t)$平移一个距离a得函数$\delta(t-a)$，它表示集中于点$t=a$的单位脉冲。显然对于任意连续函数$\varphi(t)$有

$$\int_{-\infty}^{\infty}\delta(t-a)\varphi(t)\,dt=\int_{-\infty}^{\infty}\delta(t)\varphi(t+a)dt=\varphi(a) \tag{3.2.28}$$

δ 函数的谱函数几乎“不要算”就可以得到

$$\int_{-\infty}^{\infty}\delta(t)e^{-2\pi ist}dt=e^{-2\pi ist}\Big|_{t=0}\equiv 1,\quad -\infty<s<\infty$$

$$\int_{-\infty}^{\infty}\delta(t-a)e^{-2\pi ist}dt=e^{-2\pi ist}\Big|_{t=a}=e^{-2\pi ias},\quad -\infty<s<\infty$$

因此

$$\delta(t)\sim 1,\ \delta(t-a)\sim e^{-2\pi ias} \tag{3.2.29}$$

由互逆性又得

$$1\sim\delta(s),\quad e^{2\pi iat}\sim\delta(s-a) \tag{3.2.30}$$

由此得

$$\int_{-\infty}^{\infty}e^{-2\pi ist}dt=\int_{-\infty}^{\infty}e^{2\pi ist}dt=\delta(s) \tag{3.2.31}$$

$$\int_{-\infty}^{\infty}e^{-2\pi i(s-s')t}dt=\int_{-\infty}^{\infty}e^{2\pi i(s-s')t}dt=\delta(s-s') \tag{3.2.32}$$

这就是三角函数在无穷轴 $-\infty<t<\infty$ 上的正交性。 如同正交关系 (3.1.5) 在傅氏级数里一样,正交关系(3.2.31)在傅氏积分里起着基本的作用。

通常对于两个函数 f、g 定义其内积为

$$\int_{-\infty}^{\infty}f^*(t)g(t)dt$$

在傅氏变换下,内积是不变的。当 $f(t)\sim F(s)$, $g(t)\sim G(s)$时恒有

$$\int_{-\infty}^{\infty}f^*(t)g(t)dt=\int_{-\infty}^{\infty}F^*(s)G(s)ds \tag{3.2.33}$$

取 $f=g$ 则成为巴色瓦公式

$$\int_{-\infty}^{\infty}|f(t)|^2dt=\int_{-\infty}^{\infty}|F(s)|^2ds \tag{3.2.34}$$

公式(3.2.33)可以用正交关系式(3.2.32)推导如下:

$$\begin{aligned}\int_{-\infty}^{\infty}f^*(t)g(t)dt&=\int_{-\infty}^{\infty}dt\int_{-\infty}^{\infty}[F(s)e^{2\pi ist}]^*ds\int_{-\infty}^{\infty}G(s')e^{2\pi is't}ds'\\&=\int_{-\infty}^{\infty}F^*(s)ds\int_{-\infty}^{\infty}G(s')ds'\int_{-\infty}^{\infty}e^{2\pi i(s'-s)t}dt\\&=\int_{-\infty}^{\infty}F^*(s)ds\int_{-\infty}^{\infty}G(s')\delta(s'-s)ds'=\int_{-\infty}^{\infty}F^*(s)G(s)ds\end{aligned}$$

3.2.3 广义微分

考虑台阶函数

$$Y(t)=\begin{cases}0, & t<0\\1, & t>0\end{cases} \tag{3.2.35}$$

这是典型的间断函数,在间断点 $t=0$ 处有跃值

$$\lim_{h\to 0}\left[Y\left(0+\frac{h}{2}\right)-Y\left(0-\frac{h}{2}\right)\right]=Y(+0)-Y(-0)=1-0=1$$

在 $t\neq 0$ 处 $Y'(t)=0$。但是, 如果就此认为 $Y(t)$在 $-\infty<t<\infty$ 上整体的导函数为0则是

错误的，因为导函数是表示函数的变化率的，而 $Y(t)$ 在 $t=0$ 处的突变没有得到反映。事实上

$$Y'(t)=0,\quad t\neq 0$$

$$Y'(t)\big|_{t=0}=\lim_{h\to 0}\frac{1}{h}\left[Y\left(0+\frac{h}{2}\right)-Y\left(0-\frac{h}{2}\right)\right]=\lim_{h\to 0}\frac{1}{h}[Y(+0)-Y(-0)]=\infty$$

此外，作为导函数 $Y'(t)$ 应与原函数 $Y(t)$ 有下列积分关系

$$\int_{-\infty}^{\infty}Y'(t)dt=Y(\infty)-Y(-\infty)=1-0=1$$

将这三点与式 (3.2.23、3.2.24) 作比较后，就有一切理由认为 $Y(t)$ 的导函数 $Y'(t)$ 应该是 δ 函数，即

$$Y'(t)=\delta(t) \tag{3.2.36}$$

这样才正确地反映了 $Y(t)$ 的间断性对于导函数的贡献。为了对此以及类似情况给以明确的含义，需要把导数的概念加以推广。

先回到通常的情况。设 $f(t)$ 在 $-\infty<t<\infty$ 上处处有连续的导数 $f'(t)$。任取充分光滑并在无穷远处连同其导数衰减得充分快的函数 $\varphi(t)$——以后将统称这样的函数为良函数。根据分部积分，

$$\int_{-\infty}^{\infty}f'(t)\varphi(t)dt=[f\varphi]_{-\infty}^{\infty}-\int_{-\infty}^{\infty}f(t)\varphi'(t)dt$$

由于 $f\varphi$ 在 $\pm\infty$ 处为 0，因此有

$$\int_{-\infty}^{\infty}f'(t)\varphi(t)dt=-\int_{-\infty}^{\infty}f(t)\varphi'(t)dt \tag{3.2.37}$$

对一切良函数 φ 均成立。我们将在这个基础上将导数的概念作推广。

任意函数 $f(t)$ 的广义导数就是这样的函数——记作 $f'(t)$——使得对于一切良函数 φ 而言式(3.2.37)都成立。当 $f(t)$ 在通常意义下有导数，这样定义的广义导数是与之一致的；但当在通常意义下没有导数或者有如式(3.2.23)那样的奇异性时，则这个定义有新的内容。例如取 $f(t)=Y(t)$，于是，对任意良函数 φ，

$$-\int_{-\infty}^{\infty}Y(t)\varphi'(t)dt=-\int_{0}^{\infty}\varphi'(t)dt=-\varphi(\infty)+\varphi(0)=\int_{-\infty}^{\infty}\delta(t)\varphi(t)dt$$

这就肯定了式(3.2.36)。

设 $f(t)$ 在 $x=c$ 处有间断，具有有限的跃值

$$-\infty<f(c+0)-f(c-0)<\infty$$

在 $x\neq c$ 处，具有通常意义的导数——为了与广义导数相区别，记之为 $g(t)$，试问 $f(t)$ 的广义导数 $f'(t)$ 是什么？为此，任取良函数 $\varphi(t)$，考虑到 f 的间断性：

$$-\int_{-\infty}^{\infty}f\varphi'dt=-\int_{-\infty}^{c-0}f\varphi'dt-\int_{c+0}^{\infty}f\varphi'dt=[f\varphi]_{c-0}^{-\infty}+\int_{-\infty}^{c-0}g\varphi dt+[f\varphi]_{\infty}^{c+0}+\int_{c+0}^{\infty}g\varphi dt$$

$$=\int_{-\infty}^{\infty}g\varphi dt+[f\varphi]_{c-0}^{c+0}$$

由于

$$[f\varphi]_{c-0}^{c+0}=f(c+0)\varphi(c+0)-f(c-0)\varphi(c-0)$$

$$\varphi(c-0)=\varphi(c+0)=\varphi(c)=\int_{-\infty}^{\infty}\delta(t-c)\varphi(t)dt$$

因此，对于一切良函数φ恒有

$$-\int_{-\infty}^{\infty} f(t)\varphi'(t)dt=\int_{-\infty}^{\infty}\{g(t)+[f(c+0)-f(c-0)]\delta(t-c)\}\varphi(t)dt$$

所以f的广义导数为

$$f'(t)=g(t)+[f(c+0)-f(c-0)]\delta(t-c) \tag{3.2.38}$$

当$f(t)$有多个间断点$t=c_1,\ c_2,\ \cdots,\ c_m$时，则类似地有

$$f'(t)=g(t)+\sum_{k=1}^{m}[f(c_k+0)-f(c_k-0)]\delta(t-c_k) \tag{3.2.39}$$

因此，当函数有间断时，它的广义导数是通常意义下的导数再加上间断跃值引起的脉冲项。这就正确地补足了间断性对于导数的贡献，而且导数的这种推广是合理的。注意式(3.2.39)实质上是分部积分公式。如果式中的$f'(t)$取为其通常意义下的导数，则等式可以不成立。因此运用广义导数无非就是使分部积分公式得以正确的表达。

显然(3.2.36)就是(3.2.38)，(3.2.39)的特例，据此可以得出，例如

$$(\operatorname{sign} t)'=2\delta(t) \tag{3.2.40}$$

$$(\operatorname{rect} t)'=\delta\left(t+\frac{1}{2}\right)-\delta\left(t-\frac{1}{2}\right) \tag{3.2.41}$$

当函数$f_n\to f$时，相应地广义导数必有$f_n'\to f'$。这是因为，任意取良函数$\varphi(t)$，当$f_n\to f$时必有

$$\int_{-\infty}^{\infty} f_n\varphi'\,dt\to\int_{-\infty}^{\infty} f\varphi'\,dt$$

这就是

$$\int_{-\infty}^{\infty} f_n'\varphi\,dt\to\int_{-\infty}^{\infty} f'\varphi\,dt$$

即$f_n'\to f'$。

脉冲函数$\delta(t)$也有广义导数$\delta'(t)$，对一切良函数φ恒有

$$\int_{-\infty}^{\infty}\delta'(t)\varphi(t)dt=-\int_{-\infty}^{\infty}\delta(t)\varphi'(t)dt=-\varphi'(0) \tag{3.2.42}$$

由于$\delta(t)$可以表为

$$\delta(t)=\lim_{n\to\infty} n\operatorname{rect} nt=\lim_{n\to\infty} n\Lambda(nt) \tag{3.2.43}$$

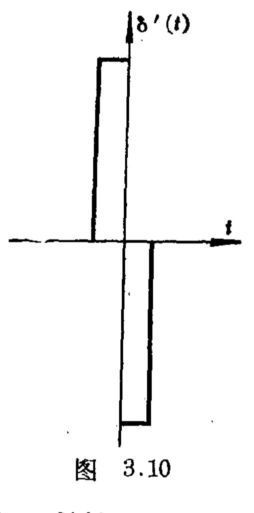
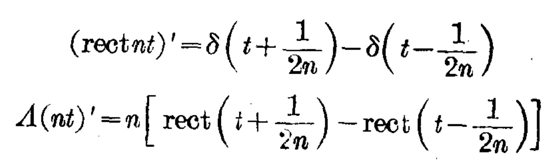


图 3.10

并且



$$(\operatorname{rect} nt)'=\delta\left(t+\frac{1}{2n}\right)-\delta\left(t-\frac{1}{2n}\right)$$

$$\Lambda(nt)'=n\left[\operatorname{rect}\left(t+\frac{1}{2n}\right)-\operatorname{rect}\left(t-\frac{1}{2n}\right)\right]$$

因此$\delta'(t)$也可以表为

$$\begin{aligned}\delta'(t)&=\lim_{n\to\infty} n\left[\delta\left(t+\frac{1}{2n}\right)-\delta\left(t-\frac{1}{2n}\right)\right]\\&=\lim_{n\to\infty} n^2\left[\operatorname{rect}\left(t+\frac{1}{2n}\right)-\operatorname{rect}\left(t-\frac{1}{2n}\right)\right]\end{aligned} \tag{3.2.44}$$

见图3.10，当n增大时，两臂愈窄而振幅愈大，趋近于$\delta'(t)$。

和$\delta(t)$相似，$\delta'(t)$也是集中于一个点$t=0$的广义函数，但比$\delta(t)$有尖锐的奇异性，它在$t\neq 0$处为0，而在$t=0$处摆动于$\pm\infty$之间。$\delta'(t)$有筛取作用，即取$x=0$处的导数值$-\varphi'(0)$。通常在物理上$\delta'(t)$代表位于$x=0$处的单位偶极子或单位矩。

广义导数 $f'(t)$ 的广义导数叫做二阶广义导数，记为 $f''(t)$，…；$(m-1)$ 阶广义导数，$f^{(m-1)}(t)$ 的广义导数为 m 阶的广义导数 $f^{(m)}(t)$，这相当于

$$\int_{-\infty}^{\infty} f^{(m)}(t)\varphi(t)dt=(-1)^m\int_{-\infty}^{\infty} f(t)\varphi^{(m)}(t)dt \tag{3.2.45}$$

对于一切良函数 φ 成立。

任取 $h>0$，对任意函数 $f(t)$ 可以定义中心差分算子 Δ_h:

$$\Delta_h f(t)=\frac{1}{h}\left[f\left(t+\frac{h}{2}\right)-f\left(t-\frac{h}{2}\right)\right] \tag{3.2.46}$$

在广义导数的意义下恒有

$$\lim_{h\to 0}\Delta_h f(t)=f'(t) \tag{3.2.47}$$

这是因为，对于任意良函数 φ，

$$\begin{aligned}\int_{-\infty}^{\infty}\Delta_h f(t)\varphi(t)dt&=\int_{-\infty}^{\infty}\frac{1}{h}\left[f\left(t+\frac{h}{2}\right)-f\left(t-\frac{h}{2}\right)\right]\varphi(t)dt\\&=\int_{-\infty}^{\infty}f(x)\frac{1}{h}\left[\varphi\left(x-\frac{h}{2}\right)-\varphi\left(x+\frac{h}{2}\right)\right]dx\end{aligned}$$

当 $h\to 0$ 时，$\frac{1}{h}\left[\varphi\left(x-\frac{h}{2}\right)-\varphi\left(x+\frac{h}{2}\right)\right]\to-\varphi'(x)$，因此

$$\int_{-\infty}^{\infty}\lim\Delta_h f(t)\varphi(t)dt=-\int_{-\infty}^{\infty}f(t)\varphi'(t)dt=\int_{-\infty}^{\infty}f'(t)\varphi(t)dt$$

反复运用差分算子 Δ_h 得高阶差分算子：

$$\Delta_h^2 f(t)=\Delta_h(\Delta_h f(t)),\ \cdots \tag{3.2.48}$$

$$\Delta_h^m f(t)=\Delta_h(\Delta_h^{m-1}f(t)) \tag{3.2.49}$$

显然有

$$\begin{aligned}\Delta_h^2 f(t)&=\Delta_h(\Delta_h f(t))=\Delta_h\frac{1}{h}\left[f\left(t+\frac{h}{2}\right)-f\left(t-\frac{h}{2}\right)\right]\\&=\frac{1}{h^2}\left[f\left(t+\frac{h}{2}+\frac{h}{2}\right)-f\left(t+\frac{h}{2}-\frac{h}{2}\right)-f\left(t-\frac{h}{2}+\frac{h}{2}\right)+f\left(t-\frac{h}{2}-\frac{h}{2}\right)\right]\\&=\frac{1}{h^2}[f(t+h)-2f(t)+f(t+h)]\end{aligned}$$

用数学归纳法不难证明

$$\Delta_h^m f(t)=\frac{1}{h^m}\sum_{k=0}^{m}(-1)^k C_k^m f\left(t+\frac{mh}{2}-kh\right),\quad C_k^m=\frac{m!}{k!(m-k)!} \tag{3.2.50}$$

并且有

$$\lim_{h\to 0}\Delta_h^m f(t)=f^{(m)}(t) \tag{3.2.51}$$

右端为 m 阶广义导数。

在谐波分析中函数及其导数的傅氏变换有极简单的关系，即表 3.1 中所列的

$$f(t)\sim F(s),\quad f'(t)\sim 2\pi isF(s) \tag{3.2.52}$$

应该指出，这样的关系只是在广义导数的意义下成立。事实上，设 $f(t)\sim F(s)$，即

$$F(s)=\int_{-\infty}^{\infty}f(t)e^{-2\pi ist}dt,$$

由分部积分公式(3.2.37)——它仅对广义导数才成立——

$$\int_{-\infty}^{\infty}f'(t)e^{-2\pi ist}dt=-\int_{-\infty}^{\infty}f(t)(e^{-2\pi ist})'dt=2\pi is\int_{-\infty}^{\infty}f(t)e^{-2\pi ist}dt$$

因此式(3.2.52)成立。

我们知道，如果 $g'(t)\equiv 0$，则 $g(t)\equiv c=$常数。根据式(3.2.52)这就等价于：如果 $2\pi isG(s)\equiv 0$，则 $G(s)=c\delta(s)$。

函数 $Y(t)$ 的谱函数可从以上的简单关系导出。事实上，命 $Y(t)\sim F(s)$，由于 $Y'(t)=\delta(t)$，得 $2\pi isF(s)\equiv 1$，因此，$2\pi is\left[F(s)-\dfrac{1}{2\pi is}\right]\equiv 0$，从而 $F(s)-\dfrac{1}{2\pi is}=c\delta(s)$，即 $F(s)=\dfrac{1}{2\pi is}+c\delta(s)$，常数 c 待定。由于 $1\equiv Y(t)+Y(-t)$，对此作傅氏变换得

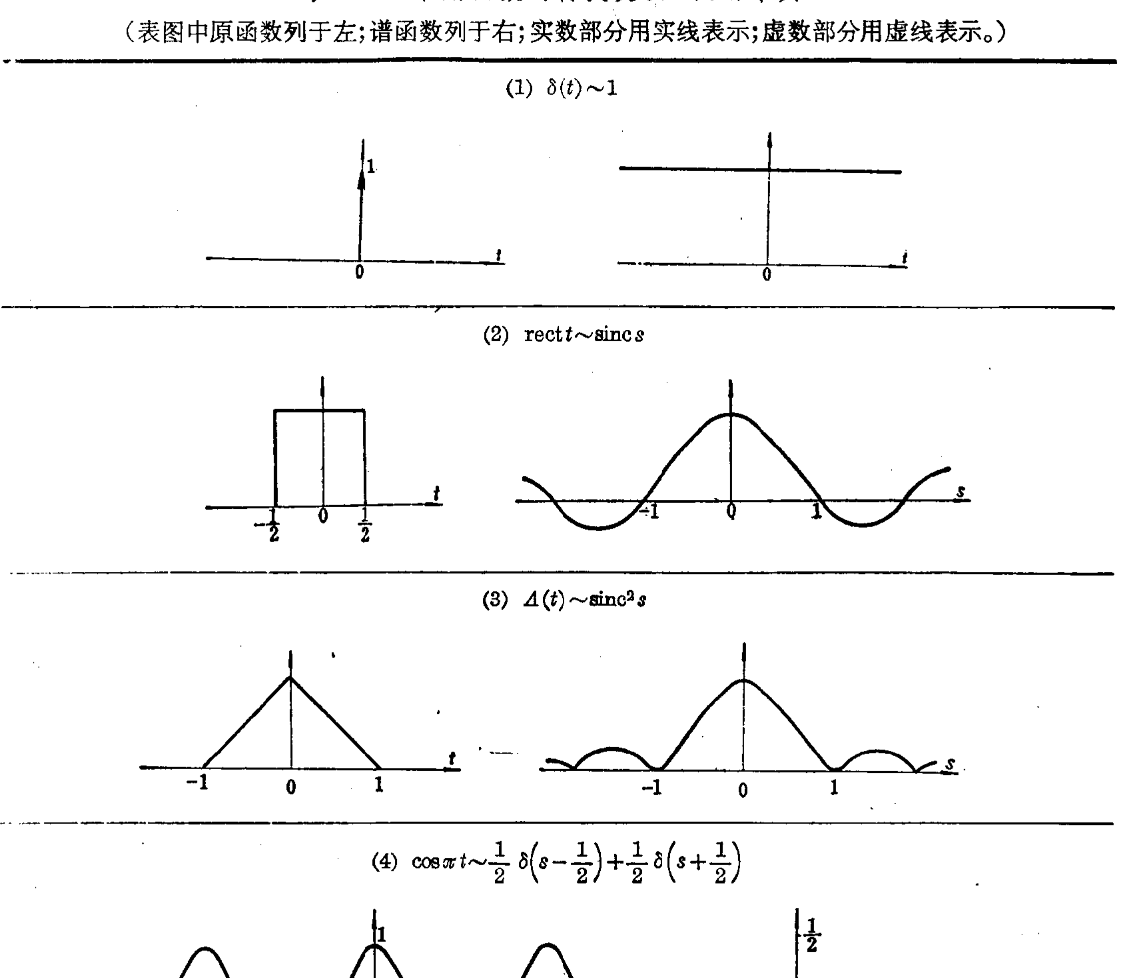
$$\delta(s)=F(s)+F(-s)=\frac{1}{2\pi is}+c\delta(s)+\frac{1}{2\pi i(-s)}+c\delta(-s)=2c\delta(s)$$ 故 $c=\dfrac{1}{2}$ 而

$$Y(t)\sim\frac{1}{2\pi is}+\frac{1}{2}\delta(s) \tag{3.2.53}$$

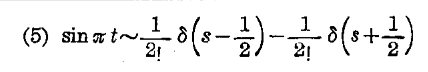
以后广义导数将统称为导数，并采用普通的记号如 f'，$f^{(p)}$，$\dfrac{df}{dt}$，$\dfrac{d^mf}{dt^m}$ 等等。

表 3.2 常用函数的傅氏变换"图解字典"

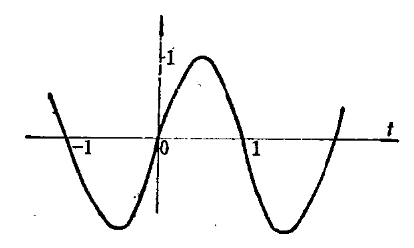
（表图中原函数列于左；谱函数列于右；实数部分用实线表示；虚数部分用虚线表示。）

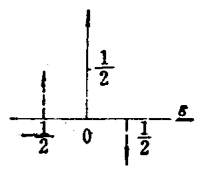


（续表）

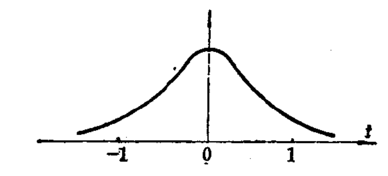


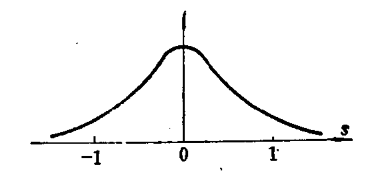
(5) $\sin \pi t \sim \frac{1}{2i}\delta\left(s-\frac{1}{2}\right)-\frac{1}{2i}\delta\left(s+\frac{1}{2}\right)$

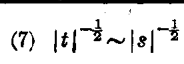




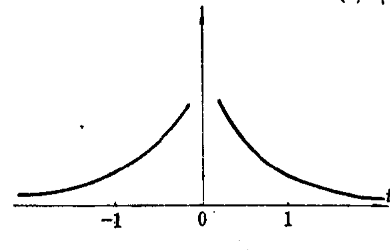
(6) $e^{-\pi t^2} \sim e^{-\pi s^2}$

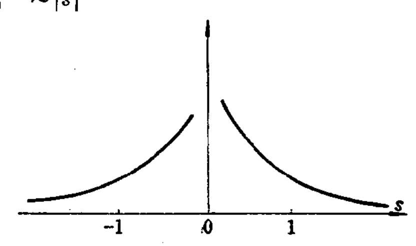




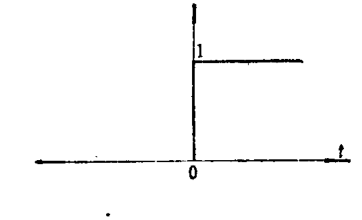


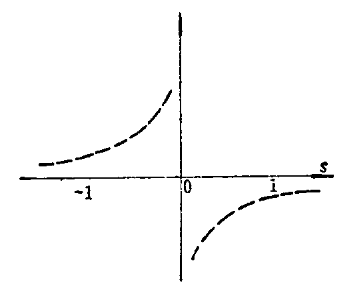
(7) $|t|^{-\frac{1}{2}} \sim |s|^{-\frac{1}{2}}$



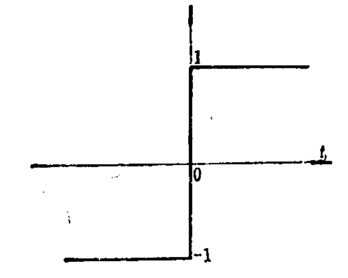


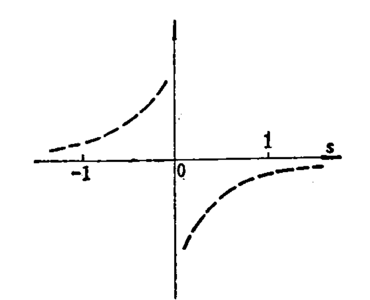
(8) $Y(t) \sim \frac{1}{2\pi i s}+\frac{1}{2}\delta(s)$





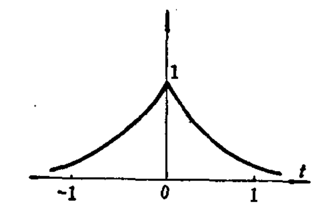
(9) $\operatorname{sign} t \sim \frac{1}{\pi i s}$

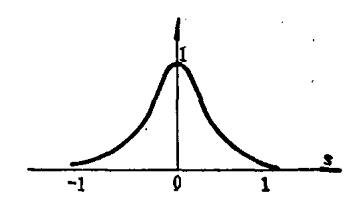




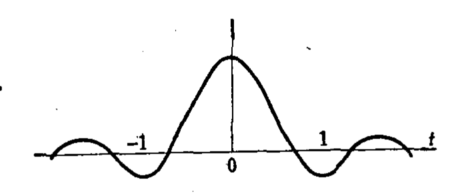
（续表）

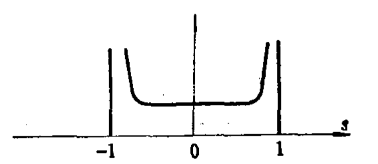
(10) $e^{-|t|} \sim \dfrac{2}{1+(2\pi s)^2}$



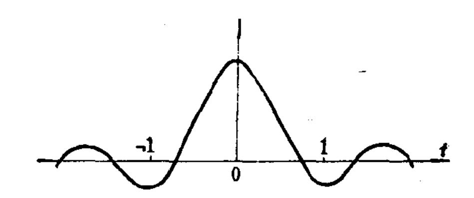


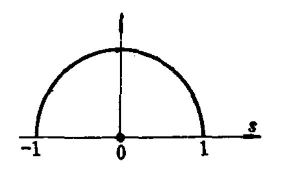
(11) $J_0(2\pi t) \sim \dfrac{1}{\pi\sqrt{1-s^2}} \operatorname{rect} s/2$



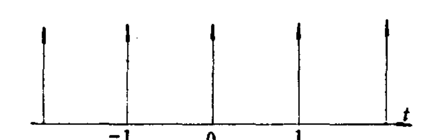


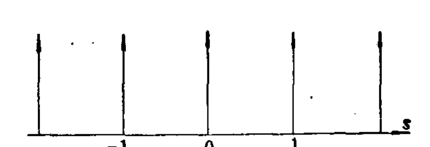
(12) $\dfrac{J_1(2\pi t)}{2t} \sim \sqrt{1-s^2} \operatorname{rect} s/2$



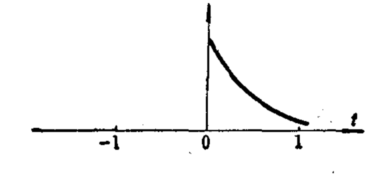


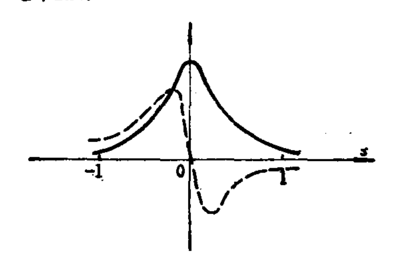
(13) $\sum_{n=-\infty}^{\infty} \delta(t-n) \sim \sum_{n=-\infty}^{\infty} \delta(s-n)$





(14) $Y(t)\, e^{-|t|} \sim \dfrac{1}{1+2\pi i s}$





§3.3 卷积与傅氏变换的对偶性质

3.3.1 卷积的定义和性质

对于两个函数

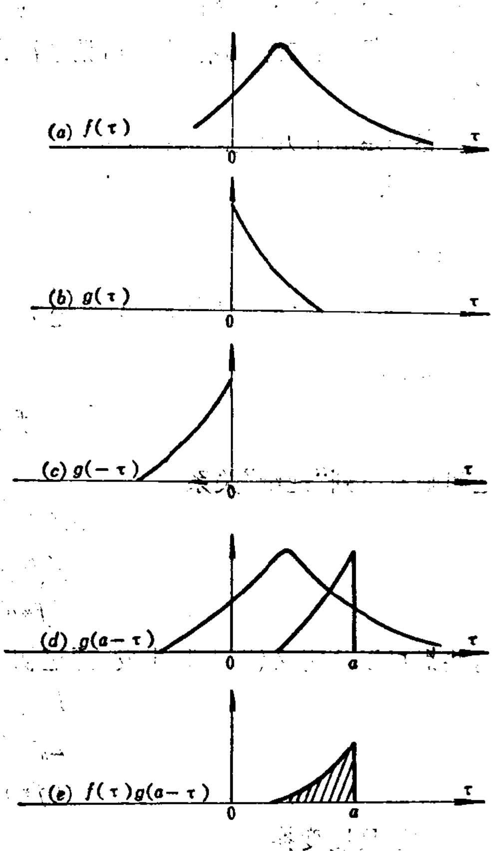
$$f(t),\ g(t),\ -\infty<t<\infty$$

定义另一个函数,称为 f 与 g 的卷积,记作 $f(t)*g(t)$ 或 $(f*g)(t)$:

$$f(t)*g(t)=\int_{-\infty}^{\infty}f(\tau)g(t-\tau)d\tau,\quad -\infty<t<\infty \tag{3.3.1}$$

这里自然默认 f 与 g 在无穷远处衰减得足够快,以使积分有意义。

图 3.11 是卷积构成的示意图。其中 (a), (b) 表示 f, g 的图形。(c) 表示 $g(\tau)$ 的反向 $g(-\tau)$。(d) 表示 g 反向再平移一个距离 a, 即 $g(-(\tau-a))=g(a-\tau)$, $t=a$ 为卷积的计算点。(e) 表示乘积 $f(\tau)g(a-\tau)$ 的图形, 阴影部分表示积分 $\int_{-\infty}^{\infty}f(\tau)g(a-\tau)d\tau$, 这就是卷积在 $t=a$ 处的值, 表为 (f) 中的竖直线的高值。



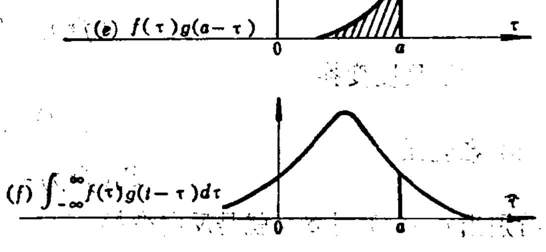


图 3.11

一个重要的事实是: 在傅氏变换之下, 函数的卷积与通常意义下的乘积是互换的, 这是傅氏变换的一种重要的对偶性质。

卷积定理: 设 $f(t)\sim F(s)$, $g(t)\sim G(s)$, 则有

$$f(t)*g(t)\sim F(s)\cdot G(s) \tag{3.3.2}$$

$$f(t)\cdot g(t)\sim F(s)*G(s) \tag{3.3.3}$$

这是很容易验证的, 例如对于第一式,

$$\int_{-\infty}^{\infty}e^{-2\pi ist}dt\int_{-\infty}^{\infty}f(\tau)g(t-\tau)d\tau$$

$$=\int_{-\infty}^{\infty}f(\tau)e^{-2\pi is\tau}d\tau\int_{-\infty}^{\infty}g(t-\tau)e^{-2\pi is(t-\tau)}dt$$

$$=F(s)\int_{-\infty}^{\infty}g(x)e^{-2\pi isx}dx=F(s)\cdot G(s)$$

卷积运算还有一些简单的规律。利用积分变量的代换可以看到

$$(f*g)(t)=\int_{-\infty}^{\infty}f(\tau)g(t-\tau)d\tau=\int_{-\infty}^{\infty}f(t-s)g(s)ds=\int_{-\infty}^{\infty}g(s)f(t-s)ds=(g*f)(t)$$

$$[(f*g)*h](t)=\int_{-\infty}^{\infty}(f*g)(s)h(t-s)ds=\int_{-\infty}^{\infty}h(t-s)ds\int_{-\infty}^{\infty}f(\tau)g(s-\tau)d\tau$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(\tau)g(s-\tau)h(t-s)ds\,d\tau$$

$$(f*(g*h))(t)=\int_{-\infty}^{\infty}f(\tau)(g*h)(t-\tau)d\tau=\int_{-\infty}^{\infty}f(\tau)d\tau\int_{-\infty}^{\infty}g(\sigma)h(t-\tau-\sigma)d\sigma$$
$$=\int_{-\infty}^{\infty}f(\tau)d\tau\int_{-\infty}^{\infty}g(s-\tau)h(t-s)ds=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(\tau)g(s-\tau)h(t-s)ds\,d\tau$$
$$=((f*g)*h)(t)$$

因此卷积服从

交换律:

$$f*g=g*f \tag{3.3.4}$$

结合律:

$$(f*g)*h=f*(g*h) \tag{3.3.5}$$

即卷积"因子"可以任意改变次序和组合而值不变。严格说来,结合律的成立是有条件的,其条件为三个函数中至少有两个是紧凑的,或者三个函数不为零的范围都从同一方向(例如都从左方或都从右方)为有界。所谓紧凑函数是指它在某个有界区间以外恒为零,见第一章§1.3。

卷积运算和几种基本运算如

平移:

$$T_a f(t)\equiv f(t-a)$$

差分:

$$\Delta_h f(t)\equiv\frac{1}{h}\left[f\left(t+\frac{h}{2}\right)-f\left(t-\frac{h}{2}\right)\right]=\frac{1}{h}(T_{-h/2}-T_{h/2})f(t) \tag{3.3.7}$$

微分:

$$f'(t)\equiv\frac{d}{dt}f(t) \tag{3.3.8}$$

之间则有下列"配给"关系

$$T_a(f*g)=T_af*g=f*T_ag \tag{3.3.9}$$
$$\Delta_h(f*g)=\Delta_hf*g=f*\Delta_hg \tag{3.3.10}$$
$$(f*g)'=f'*g=f*g' \tag{3.3.11}$$

这些都可以从定义(3.3.1)出发,通过初等的演算加以验证。由此进一步又有

$$T_{a+b}(f*g)=T_af*T_bg \tag{3.3.12}$$
$$\Delta_h^{p+q}(f*g)=\Delta_h^pf*\Delta_h^qg \tag{3.3.13}$$
$$(f*g)^{(p+q)}=f^{(p)}*g^{(q)} \tag{3.3.14}$$

对于尺度变换

$$S_af(t)\equiv f(at),\quad a\neq0 \tag{3.3.15}$$

则不难验证

$$S_a(f*g)=|a|(S_af*S_ag) \tag{3.3.16}$$

特别当 $a=-1$ 时 S_{-1} 就是反向变换

$$S_{-1}f(t)=f(-t) \tag{3.3.17}$$
$$S_{-1}(f*g)=S_{-1}f*S_{-1}g \tag{3.3.18}$$

注意:当 f 为偶函数时,$S_{-1}f=f$;当 f 为奇函数时,$S_{-1}f=-f$。于是可知两个偶函数或两个奇函数的卷积是偶函数,奇函数与偶函数的卷积则是奇函数。

此外,卷积 $f*g$ 在 $(-\infty,\infty)$ 上的积分值等于 f 与 g 的积分值的乘积,即*

$$\int_{-\infty}^{\infty}(f(t)*g(t))dt=\int_{-\infty}^{\infty}f(t)dt\cdot\int_{-\infty}^{\infty}g(t)dt \tag{3.3.19}$$

这是因为

$$\int_{-\infty}^{\infty}dt\int_{-\infty}^{\infty}f(t-\tau)g(\tau)d\tau=\int_{-\infty}^{\infty}g(\tau)d\tau\int_{-\infty}^{\infty}f(t-\tau)dt=\int_{-\infty}^{\infty}g(\tau)d\tau\int_{-\infty}^{\infty}f(t)dt$$

我们将通过一些例子来说明卷积的作用。

首先取台阶函数

$$Y(t)=\begin{cases}0, & t<0\\ 1, & t>0\end{cases} \tag{3.3.20}$$

作为卷积因子,

$$Y(t)*f(t)=\int_{-\infty}^{\infty}f(\tau)Y(t-\tau)d\tau=\int_{-\infty}^{t}f(\tau)d\tau \tag{3.3.21}$$

因此,用台阶函数 Y 卷乘 f 相当于作 f 的不定积分,起着积分算子的作用,从而把 f 的光滑度提高一阶,并且有

$$\frac{d}{dt}[Y(t)*f(t)]=f(t) \tag{3.3.22}$$

如果更取 $f=Y$,则有

$$Y(t)*Y(t)=\int_{-\infty}^{t}Y(\tau)d\tau=\begin{cases}0, & t<0\\ \int_0^t d\tau=t, & t>0\end{cases} \tag{3.3.23}$$

$$\frac{d}{dt}[Y(t)*Y(t)]=Y(t) \tag{3.3.24}$$

因此,$Y*Y$ 已经是连续函数,一阶导数在 $t=0$ 处有间断。

取矩形函数

$$\mathrm{rect}\,t=\begin{cases}0, & |t|>\dfrac{1}{2}\\ 1, & |t|<\dfrac{1}{2}\end{cases} \tag{3.3.25}$$

作为卷积因子则有

$$\mathrm{rect}\,t*f(t)=\int_{-\infty}^{\infty}f(\tau)\mathrm{rect}\,(t-\tau)d\tau=\int_{t-\frac{1}{2}}^{t+\frac{1}{2}}f(\tau)d\tau \tag{3.3.26}$$

这相当于对 f 作积分中值函数,用 f 在区间$\left[t-\frac{1}{2},\ t+\frac{1}{2}\right]$上的积分平均值以代 $f(t)$,并且也起平滑化的作用,提高一阶光滑度。例如取 $f(t)=\mathrm{rect}\,t$,则通过初等演算可得

$$\mathrm{rect}\,t*\mathrm{rect}\,t=\int_{t-\frac{1}{2}}^{t+\frac{1}{2}}\mathrm{rect}\,\tau\,d\tau=\begin{cases}0, & t\leqslant -1\\ 1+t, & -1\leqslant t\leqslant 0\\ 1-t, & 0\leqslant t\leqslant 1\\ 0, & t\geqslant 1\end{cases} \tag{3.3.27}$$

这就是三角形函数(§3.2),

$$\mathrm{rect}\,t*\mathrm{rect}\,t=\Lambda(t)=\begin{cases}0, & |t|\geqslant 1\\ 1-|t|, & |t|\leqslant 1\end{cases} \tag{3.3.28}$$

这是连续函数,一阶导数在 $t=0$, ± 1 有间断。

取三角形函数 $\Lambda(t)$ 为卷积因子时,

$$\Lambda(t)*f(t)=\int_{-\infty}^{\infty}f(\tau)\Lambda(t-\tau)d\tau=\int_{t-1}^{t+1}f(\tau)(1-|\tau-t|)d\tau \tag{3.3.29}$$

相当于用 f 在区间 $[t-1,\ t+1]$ 的加权积分平均值以代 $f(t)$，所加的权作三角形分布。上面用间断函数 Y 或 rect 作卷乘时提高一阶光滑度。这里 Λ 是连续函数，用 Λ 卷乘相当于用 rect 卷乘两次，故光滑度提高两阶。

一般说来，用通常的函数 g 卷乘于 f 时总是起平滑化的作用，即积分的作用。g 的光滑度愈高，则 $g*f$ 相对于 f 光滑度提高愈多。

反之当取 g 为奇异函数以作卷积，则情况不同。例如取为脉冲函数 δ 或其导数 δ'，则有

$$\delta*f=\int_{-\infty}^{\infty}\delta(\tau)f(t-\tau)d\tau=f(t)$$

$$\delta'*f=\int_{-\infty}^{\infty}\delta'(\tau)f(t-\tau)d\tau=-\int_{-\infty}^{\infty}\delta(\tau)\frac{d}{d\tau}f(t-\tau)d\tau=\int_{-\infty}^{\infty}\delta(\tau)f'(t-\tau)d\tau=f'(t)$$

即

$$\delta*f=f,\quad \delta'*f=f',\quad \delta^{(m)}*f=f^{(m)} \tag{3.3.30}$$

因此，用 δ 函数卷乘任意 f 时使 f 不变，而用 δ' 卷乘则相当于作微分。一般说来，用一个奇异程高于 $\delta(t)$ 的函数 $g(t)$ 作卷乘时起着降低光滑度的作用，即微分的作用，其程度随 $g(t)$ 的奇异度的提高而愈甚。

3.3.2 样条函数及其傅氏变换

分段多项式包括样条函数在函数插值和曲线拟合中是一项重要的工具(见第一章)。这里将讨论一些基本样条的差分表达式和卷积表达式，以及它们的谱函数。

当函数能表成分段次数 $\leqslant m$ 的多项式，并在分段点即节点上直至 $m-1$ 阶导数为连续时，叫做 m 次样条。0 次样条就是分段常数。

函数 Y 及 $Y*Y$ 分别是 0 次及一次样条，它们的支集都是 $[0,\ \infty)$，是无限的，但从左方为有界。函数 rect, $\Lambda=\text{rect}*\text{rect}$ 也分别是 0 次及一次样条，它们的支集 $\left[-\frac{1}{2},\frac{1}{2}\right]$, $[-1,1]$ 是有限的。可以把这两类样条推广到高次。

显然，可以取“截断幂”

$$t_+^m=\begin{cases}0, & t<0\\ t^m, & t>0\end{cases} \tag{3.3.31}$$

这是分两段的 m 次多项式，分段点是 $t=0$，Y 及 $Y*Y$ 就是 $m=0$ 及 1 的特例。为了以后的方便，引进一个非本质的常因子，对于 $p=1, 2, 3, \cdots$，命

$$Y_p(t)=\frac{1}{(p-1)!}t_+^{p-1}=\begin{cases}0, & t<0\\ \dfrac{1}{(p-1)!}t^{p-1}, & t>0,\ 0!=1\end{cases} \tag{3.3.32}$$

这是以 $t=0$ 为节点的分段 $p-1$ 次多项式。为了方便，约定

$$Y_0(t)=\delta(t) \tag{3.3.33}$$

并且有

$$Y_1(t)=Y(t) \tag{3.3.34}$$

$$Y_2(t)=Y(t)*Y(t) \tag{3.3.35}$$

显然可见

$$Y_p'(t)=Y_{p-1}(t) \tag{3.3.36}$$

由此递推出

$$Y_p^{(q)}(t)=Y_{p-q}(t),\ 0\leqslant q\leqslant p \tag{3.3.37}$$

$$Y_p^{(p-1)}(t)=Y_1(t)=Y(t),\ Y_p^{(p)}(t)=Y_0(t)=\delta(t) \tag{3.3.38}$$

Y_p 直至 $p-2$ 阶导数为连续，$p-1$ 阶导数在 $t=0$ 处有间断，因此 Y_p 是 $p-1$ 次样条。另一方面

$$Y_1(t)*Y_{p-1}(t)=Y(t)*Y_{p-1}(t)$$

$$=\int_{-\infty}^{t}Y_{p-1}(\tau)d\tau=\begin{cases}0,\ t<0\\ \displaystyle\int_0^t\frac{1}{(p-2)!}\tau^{p-2}d\tau=\frac{1}{(p-1)!}t^{p-1},\ t>0\end{cases}$$

因此

$$Y_p=Y_1*Y_{p-1}=Y*Y_{p-1} \tag{3.3.39}$$

由此递推得到样条 Y_p 的卷积表达式

$$Y_p=Y_1*Y_1*\cdots*Y_1=Y*Y*\cdots*Y\quad(p\text{ 个}) \tag{3.3.40}$$

$$Y_{p+q}=Y_p*Y_q \tag{3.3.41}$$

样条 Y_p 不是紧凑函数，它不为零的范围 $(0,\ \infty)$，是无界的。但是，对于 Y_p 作差分可以得到紧凑的样条，后者由于其稳定性，在计算实践中特别重要(见第一章 1.3.5 节)。为此目的，取差分算子 Δ_h，$h=1$，简记为 Δ:

$$\begin{cases}\Delta f(t)=\Delta^1 f(t)=f\left(t+\dfrac{1}{2}\right)-f\left(t-\dfrac{1}{2}\right)\\ \Delta^2 f(t)=f(t+1)-2f(t)+f(t-1)\\ \Delta^p f(t)=\displaystyle\sum_{k=0}^{p}(-1)^k C_k^p f\left(t+\frac{p}{2}-k\right),\ C_k^p=\frac{p!}{k!(p-k)!}\end{cases} \tag{3.3.42}$$

据此，对于 $p=1,\ 2,\ 3,\ \cdots$，命(见第一章 1.5.1 节)

$$M_p(t)=\Delta^p Y_p(t)=\frac{1}{(p-1)!}\Delta^p t_+^{p-1} \tag{3.3.43}$$

为了方便，约定

$$M_0(t)=\delta(t) \tag{3.3.44}$$

由于 Y_p 为 $p-1$ 次样条，在 $t=0$ 处 $p-1$ 阶导数间断。M_p 是由 Y_p 平移迭加而得，因此也是 $p-1$ 次样条。但以

$$t=-\frac{p}{2}+k,\ k=0,\ 1,\ \cdots,\ p$$

即

$$t=-\frac{p}{2},\ -\frac{p}{2}+1,\ \cdots,\ \frac{p}{2}-1,\ \frac{p}{2} \tag{3.3.45}$$

为节点，在这些点 $p-1$ 阶导数间断。应该注意的是，当 p 为偶数时节点为整点，当 p 为奇数时则为半点。

不难算出，例如

$$M_1(t)=\Delta^1 Y_1(t)=\Delta Y(t)=\left(t+\frac{1}{2}\right)_+^0-\left(t-\frac{1}{2}\right)_+^0=\begin{cases}0,\ t<-\dfrac{1}{2}\\ 1,\ -\dfrac{1}{2}<t<\dfrac{1}{2}\\ 0,\ t>\dfrac{1}{2}\end{cases} \tag{3.3.46}$$

$$M_2(t)=\Delta^2Y_2(t)=(t+1)_+-2t_++(t-1)_+=\begin{cases}0, & t<-1\\ 1+t, & -1<t<0\\ 1-t, & 0<t<1\\ 0, & t>1\end{cases} \tag{3.3.47}$$

因此

$$M_1(t)=\operatorname{rect} t \tag{3.3.48}$$

$$M_2(t)=\Lambda(t)=\operatorname{rect} t*\operatorname{rect} t=M_1(t)*M_1(t) \tag{3.3.49}$$

更一般些，由于 (3.3.39)，(3.3.40)，(3.3.41) 以及差分算子对于卷积的“配给”关系 (3.3.13)

$$M_p=\Delta^pY_p=\Delta^p(Y_1*Y_{p-1})=\Delta^1Y_1*\Delta^{p-1}Y_{p-1}=M_1*M_{p-1}$$

即

$$M_p=M_1*M_{p-1}=\operatorname{rect} t*M_{p-1} \tag{3.3.50}$$

由此递推，得到样条 M_p 的卷积表达式，

$$M_p=M_1*M_1*\cdots*M_1=\operatorname{rect} t*\operatorname{rect} t*\cdots*\operatorname{rect} t \quad (p\text{ 个}) \tag{3.3.51}$$

$$M_{p+q}=M_p*M_q \tag{3.3.52}$$

注意，上述二式与式(3.3.40)(3.3.41)完全相似。

至于微分，

$$M_1'(t)=\operatorname{rect}' t=\delta\left(t+\frac{1}{2}\right)-\delta\left(t-\frac{1}{2}\right)=\Delta\delta=\Delta M_0(t) \tag{3.3.53}$$

注意到对于任意函数 $f(t)$，有

$$\left[\delta\left(t+\frac{1}{2}\right)-\delta\left(t-\frac{1}{2}\right)\right]*f(t)=f\left(t+\frac{1}{2}\right)-f\left(t-\frac{1}{2}\right)=\Delta f(t)$$

于是，根据(3.3.50)，(3.3.11)得

$$M_p'=M_1'*M_{p-1}=\left[\delta\left(t+\frac{1}{2}\right)-\delta\left(t-\frac{1}{2}\right)\right]*M_{p-1}=\Delta M_{p-1}$$

即

$$M_p'=\Delta M_{p-1} \tag{3.3.54}$$

由此递推

$$M_p^{(q)}(t)=\Delta^qM_{p-q}(t),\quad 0\leqslant q\leqslant p \tag{3.3.55}$$

$$M_p^{(p-1)}(t)=\Delta^{p-1}M_1(t)=\Delta^{p-1}\operatorname{rect} t,\quad M_p^{(p)}(t)=\Delta^pM_0(t)=\Delta^p\delta(t) \tag{3.3.56}$$

类似于公式(3.3.36～3.3.38)。

根据(3.3.26)，公式(3.3.50)可以表为递推的积分公式

$$M_p(t)=\int_{t-\frac{1}{2}}^{t+\frac{1}{2}}M_{p-1}(\tau)d\tau,\quad p=2,\ 3,\ 4,\ \cdots \tag{3.3.57}$$

从 $M_1=\operatorname{rect}$，$M_2=\Lambda$ 的性状出发，根据这个积分公式可以逐步推出 M_p 的性状：

(1) 由于 $M_1(t)$是偶函数，因此它的逐次卷积 M_2，$M_3\cdots$ 也都是偶函数

$$M_p(t)=M_p(-t) \tag{3.3.58}$$

(2) 由于 $M_1(t)$ 在 $|t|\geqslant\frac{1}{2}$ 上恒为 0，每积分一次，函数恒为 0 的范围向左右方各退 $\frac{1}{2}$，因此 M_p 是紧凑的：

$$M_p(t)\equiv 0,\quad |t|\geqslant\frac{p}{2} \tag{3.3.59}$$

(3) 由于在 $|t|<\frac{1}{2}$ 上 $M_1(t)>0$，每积分一次，函数 >0 的范围向左右方各伸 $\frac{1}{2}$，因此

$$M_p(t)>0, \quad |t|<\frac{p}{2} \tag{3.3.60}$$

(4) $M_1=\text{rect}$ 的图形是平顶的，$M_2=\Lambda$ 以 $t=0$ 为唯一的极大点。这个性质在逐次积分都被保持，即

$$M_p(0)>M_p(t), \quad t\neq 0, \quad p=2, 3, \cdots \tag{3.3.61}$$

(5) M_1 在 $(-\infty, \infty)$ 上积分值为 1，逐次以 M_1 作卷积后由于(3.3.19)，积分值不变，即

$$\int_{-\infty}^{\infty} M_p(t)dt=1, \quad p=1, 2, \cdots$$

样条 M_p 除了有积分递推公式(3.3.57)外，还有代数递推公式

$$M_p(t)=\frac{1}{p-1}\left[\left(\frac{p}{2}+t\right)M_{p-1}\left(t+\frac{1}{2}\right)+\left(\frac{p}{2}-t\right)M_{p-1}\left(t-\frac{1}{2}\right)\right] \tag{3.3.62}$$

这是导源于差分算子 Δ 的"乘积公式"

$$\Delta[f(t)g(t)]=[\Delta f(t)]g\left(t+\frac{1}{2}\right)+f\left(t-\frac{1}{2}\right)\Delta g(t) \tag{3.3.63}$$

只需把两端展开比较就得验证。取 $f(t)=t$，由于 $\Delta t=1$ 得到

$$\Delta[tg(t)]=g\left(t+\frac{1}{2}\right)+\left(t-\frac{1}{2}\right)\Delta g(t) \tag{3.3.64}$$

再将 Δ 作用于两端，又由于 $\Delta\left(t-\frac{1}{2}\right)=1$，由乘积公式又得

$$\begin{aligned}\Delta^2[tg(t)]&=\Delta g\left(t+\frac{1}{2}\right)+\Delta g\left(t+\frac{1}{2}\right)+\left(t-\frac{1}{2}-\frac{1}{2}\right)\Delta^2 g(t)\\&=2\Delta g\left(t+\frac{1}{2}\right)+\left(t-\frac{2}{2}\right)\Delta^2 g(t)\end{aligned}$$

依次类推，得到

$$\Delta^3[tg(t)]=3\Delta^2 g\left(t+\frac{1}{2}\right)+\left(t-\frac{3}{2}\right)\Delta^3 g(t)$$

$$\cdots\cdots\cdots\cdots$$

$$\Delta^p[tg(t)]=p\Delta^{p-1}g\left(t+\frac{1}{2}\right)+\left(t-\frac{p}{2}\right)\Delta^p g(t) \tag{3.3.65}$$

将此运用于

$$M_p(t)=\Delta^p Y_p(t), \quad Y_p(t)=\frac{1}{p-1}[tY_{p-1}(t)]$$

便有

$$\begin{aligned}M_p(t)&=\frac{1}{p-1}\left[p\Delta^{p-1}Y_{p-1}\left(t+\frac{1}{2}\right)+\left(t-\frac{p}{2}\right)\Delta^p Y_{p-1}(t)\right]\\&=\frac{1}{p-1}\left[pM_{p-1}\left(t+\frac{1}{2}\right)+\left(t-\frac{p}{2}\right)\Delta M_{p-1}(t)\right]\\&=\frac{1}{p-1}\left[pM_{p-1}\left(t+\frac{1}{2}\right)+\left(t-\frac{p}{2}\right)\left(M_{p-1}\left(t+\frac{1}{2}\right)-M_{p-1}\left(t-\frac{1}{2}\right)\right)\right]\end{aligned}$$

这就是(3.3.62)。

当 p 很高时，用差分公式(3.3.43)计算 M_p 是不稳定的(见第一章 §1.4)，但用递推公式(3.3.62)来计算则相对地稳定。这是因为，在 M_p 的非零范围，即 $-\frac{p}{2}<t<\frac{p}{2}$ 内恒有

$$\frac{p}{2}+t>0,\quad \frac{p}{2}-t>0,\quad \frac{1}{p-1}\left[\left(\frac{p}{2}+t\right)+\left(\frac{p}{2}-t\right)\right]=\frac{p}{p-1}$$

因此，如果对于 $M_1(t)$ 有初始的最大绝对误差 δ_1，则对于 $M_p(t)$ 的最大绝对误差 δ_p 必满足

$$\delta_p\leqslant\frac{1}{p-1}\left[\left|\frac{p}{2}+t\right|+\left|\frac{p}{2}-t\right|\right]\delta_{p-1}=\frac{p}{p-1}\delta_{p-1}$$

$$\delta_p\leqslant\frac{p}{p-1}\delta_{p-1}\leqslant\frac{p}{p-1}\cdot\frac{p-1}{p-2}\delta_{p-2}=\frac{p}{p-2}\delta_{p-2}\leqslant\cdots\leqslant p\delta_1$$

因此 δ_p 不作恶性增长。递推过程基本稳定。此外，上面所说的 M_p 的若干基本性质也可以从这个递推公式导出。

从几何上看，M_p 是单峰式的对称山丘形函数。当 p 增大时，光滑度逐步提高，"基底"逐步加宽，面积保持为 1。M_1、M_2、M_3、M_4 的曲线见第一章图 1.32。下面给出它们的分段表达式。除了通常的幂次表达外，还给出用三角形函数 $\Lambda(x)$ 的逐次平移的表达式，比较紧凑些。为了简化，采用记号

$$\lambda_i=\lambda_i(t)=\Lambda(t-i),\quad i=0,\ \pm1,\ \pm2,\ \cdots \tag{3.3.66}$$

这就是以整数 $t=0,\ \pm1,\ \pm2,\ \cdots$ 为节点的分段线性插值基函数(见第一章 1.3.5 节)。

$$\lambda_{i+\frac{1}{2}}=\lambda_{i+\frac{1}{2}}(t)=\Lambda\left(t-\left(i+\frac{1}{2}\right)\right),\quad i=0,\ \pm1,\ \pm2,\ \cdots \tag{3.3.67}$$

这就是以半整数 $t=\pm\frac{1}{2},\ \pm\frac{3}{2},\ \pm\frac{5}{2},\ \cdots$ 为节点的分段线性插值基函数。

$$M_1(t)=\left(t+\frac{1}{2}\right)_+^0-\left(t-\frac{1}{2}\right)_+^0=\begin{cases}0, & t<-\frac{1}{2}\\ 1, & -\frac{1}{2}<t<\frac{1}{2}\\ 0, & \frac{1}{2}<t\end{cases}=\operatorname{rect} t \tag{3.3.67'}$$

$$M_2(t)=(t+1)_+-2(t)_++(t-1)_+=\begin{cases}0, & t\leqslant-1\\ 1+t, & -1\leqslant t\leqslant0\\ 1-t, & 0\leqslant t\leqslant1\\ 0, & 1\leqslant t\end{cases}=\Lambda(t)=\lambda_0 \tag{3.3.68}$$

$$M_3(t)=\frac{1}{2}\left[\left(t+\frac{3}{2}\right)_+^2-3\left(t+\frac{1}{2}\right)_+^2+3\left(t-\frac{1}{2}\right)_+^2-\left(t-\frac{3}{2}\right)_+^2\right]$$

$$=\frac{1}{2}\cdot\begin{cases}0, & t\leqslant-\frac{3}{2}\\ \left(\frac{3}{2}+t\right)^2, & -\frac{3}{2}\leqslant t\leqslant-\frac{1}{2}\\ \frac{6}{4}-2t^2, & -\frac{1}{2}\leqslant t\leqslant\frac{1}{2}\\ \left(\frac{3}{2}-t\right)^2, & \frac{1}{2}\leqslant t\leqslant\frac{3}{2}\\ 0, & \frac{3}{2}\leqslant t\end{cases}=\frac{1}{2}\left[\lambda^2_{-\frac{1}{2}}+4\lambda_{-\frac{1}{2}}\lambda_{\frac{1}{2}}+\lambda^2_{\frac{1}{2}}\right] \tag{3.3.69}$$

$$M_4(t)=\frac{1}{6}[(t+2)_+^3-4(t+1)_+^3+6(t)_+^3-4(t-1)_+^3+(t-2)_+^3]$$

$$=\frac{1}{6}\cdot\begin{cases}0, & t\leqslant -2\\ (2+t)^3, & -2\leqslant t\leqslant -1\\ (2+t)^3-4(1+t)^3, & -1\leqslant t\leqslant 0\\ (2-t)^3-4(1-t)^3, & 0\leqslant t\leqslant 1\\ (2-t)^3, & 1\leqslant t\leqslant 2\\ 0, & 2\leqslant t\end{cases}$$

$$=\frac{1}{6}[\lambda_{-1}^3+6\lambda_{-1}^2\lambda_0+12\lambda_{-1}\lambda_0^2+4\lambda_0^3+12\lambda_0^2\lambda_1+6\lambda_0\lambda_1^2+\lambda_1^3] \tag{3.3.70}$$

样条函数 $M_p(t)$ 在谐波分析中有典型的意义。已知 M_1 的谱函数

$$M_1(t)=\operatorname{rect} t\sim \operatorname{sinc} s=\frac{\sin \pi s}{\pi s} \tag{3.3.71}$$

于是根据 M_p 的卷积表达式以及卷积与乘积的互换定理可知

$$M_p(t)\sim(\operatorname{sinc} s)^p=\left(\frac{\sin \pi s}{\pi s}\right)^p \tag{3.3.72}$$

随着 p 的增大，M_p 的光滑度增高，而其谱函数在 $|s|\to\infty$ 处的衰减率也增高。更确切地说，M_p 为 $p-1$ 次样条，它的 $p-2$ 阶导数连续，$p-1$ 阶导数有跳跃性间断，即有如台阶函数状，p 阶导数有脉冲式的奇异性，即有如 δ 函数，而相应的谱函数 $F(s)$ 在无穷远处按 s 的 p 次幂衰减，即 $F(s)=O(|s|^{-p})$，$|s|\to\infty$。这一事实可以推广到一般，即在傅氏变换下，光滑度与衰减率是互换的。样条 M_p 具体地反映了这一规律性。

3.3.3 卷积的物理意义

通常的物理器件，例如滤波器、光学仪器等等，所实现的功能都可以示意地表示为

输入 $f(t)\to$ [器　件] $\to$ 输出 $h(t)$

输出取决于输入以及器件自身的功能特征。在许多场合下，器件具有下列性质

(一)线性：如果　　$f_1(t),\ f_2(t)\to h_1(t),\ h_2(t)$

　　　　　则　　　$\alpha_1 f_1(t)+\alpha_2 f_2(t)\to \alpha_1 h_1(t)+\alpha_2 h_2(t)$

(二)平移不变性：如果 $f(t)\to h(t)$，则 $f(t-\tau)\to h(t-\tau)$。这时，设输入为 $\delta(t)$，输出为 $g(t)$，则 $g(t)$ 叫做器件的脉冲响应函数。由性质(二)可知，当输入为 $\delta(t-\tau)$ 时，输出为 $g(t-\tau)$。根据式(3.2.28)，任意输入 $f(t)$ 可以表为 $\delta(t-\tau)$ 的线性迭加

$$f(t)=\int_{-\infty}^{\infty} f(\tau)\delta(t-\tau)d\tau,\quad -\infty<t<\infty$$

故由性质(一)可知，相应的输出就是

$$h(t)=\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau=f(t)*g(t) \tag{3.3.73}$$

即输出函数表为输入函数与器件的特征函数(即脉冲响应函数)的卷积。

在时间序列分析中，对于函数 $f(t)$，$-\infty<t<\infty$ 定义其自相关函数

$$f_A(t)=\int_{-\infty}^{\infty} f^*(\tau)f(t+\tau)d\tau \tag{3.3.74}$$

它刻画了序列$f(t)$的自身的统计相关性。很容易看出，它就是$f(t)$与$f^*(-t)$的卷积，即

$$f_A(t)=f(t)*f^*(-t) \tag{3.3.75}$$

设$f(t)\sim F(s)$，从卷积与乘积的互换性就得到

$$f_A(t)\sim|F(s)|^2 \tag{3.3.76}$$

这就是所谓维纳-辛钦(Wiener-Хинчин)定理。$|F(s)|^2$称为函数$f(t)$的功率谱，它在许多科学技术领域里有应用。

3.3.4 傅氏变换的对偶关系

在傅氏变换下，原函数与谱函数之间有许多互换关系，即对偶关系。回顾基本性质简表(表3.1)就可看到，几乎其中每一项表示一种对偶关系。掌握这种对偶关系对于运用谐波分析的工具来解实际问题是有益的。特别重要的有光滑度与增衰率之间的互换以及“窄”与“宽”之间的互换，下面对此稍加说明：

依次考虑

$$\delta(t)\sim 1,\quad \operatorname{rect} t\sim\frac{\sin\pi s}{\pi s},\quad \Lambda(t)\sim\left(\frac{\sin\pi s}{\pi s}\right)^2$$

δ函数具有点脉冲式的奇异性，它的谱函数$\equiv 1$。$\operatorname{rect} t$为台阶状间断函数，导数$\operatorname{rect}' t=\delta\left(t+\frac{1}{2}\right)-\delta\left(t-\frac{1}{2}\right)$为点脉冲式，因此光滑度提高一阶，而谱函数在无穷处量级为$O(|s|^{-1})$，按一次幂衰减。$\Lambda(t)$本身是连续函数，一阶导数$\Lambda'(t)=\operatorname{rect}\left(t+\frac{1}{2}\right)-\operatorname{rect}\left(t-\frac{1}{2}\right)$有台阶状间断，二阶导数$\Lambda''(t)=\delta(t+1)-2\delta(t)+\delta(t-1)$为点脉冲式，光滑度又高一阶，而谱函数在无穷处衰减加快一阶，量级为$O(|s|^{-2})$。

一般地，对于山丘形样条M_p ($p=0, 1, 2$时就是δ, rect, Λ)有

$$M_p(t)\sim\left(\frac{\sin\pi s}{\pi s}\right)^p$$

$$M_p^{(p)}(t)=\Delta^p\delta(t)=\sum_{k=0}^{p}(-1)^k C_p^k\delta\left(t+\frac{p}{2}-k\right)$$

$$\left(\frac{\sin\pi s}{\pi s}\right)^p=O(|s|^{-p}),\quad |s|\to\infty$$

M_p开始在p阶导数含有点脉冲式的奇异性——即$p-1$阶导数开始含台阶状间断，$p-2$阶及以上的导数为连续。谱函数则在无穷处按p次幂衰减。这样，函数的光滑度递增对应于谱函数衰减率加快。

从另一方向也可以考虑

$$\delta(t)\sim 1,\quad \delta'(t)\sim 2\pi is,\quad \cdots,\quad \delta^{(p)}\sim(2\pi is)^p$$

光滑度递降而谱函数在无穷处的增长率递升，这里$\delta(t)$是$\delta^{(p)}(t)$的p次积分，由微分与积分的互逆，也不妨认为$\delta(t)$是$\delta^{(p)}(t)$的$(-p)$阶微分。

以上的特例反映了一个普遍的性质，即函数的光滑度与增衰率之间存在着对偶的关系。当函数f开始在p阶导数出现点脉冲式的奇异性时，对应于谱函数F在无穷处有$F(s)=O(|s|^{-p})$。这样，根据原函数的光滑度[增衰率]可以对谱函数的增衰率[光滑度]作出估计，在实践上很有用处。

函数的光滑性，即可微分性是函数逐点的性质，是一种局部的性质，而函数的增衰率，即

无穷处的渐近行为，则是一种大范围的、整体的性质。因此上述互换关系可以看为在傅氏变换下的一种局部性质与整体性质之间的对偶关系。

至于所谓宽与窄的对偶关系，在自然界的种种波动现象中都会遇到。例如，在声学领域，音响的持续时间愈短促，则音调成分愈杂，即频谱愈宽；反之，音调最纯的单音在时间上就是无限重复的简谐振动。在电子技术里，脉冲的宽度愈窄则频带愈宽。对于光的衍射，当照明孔径愈窄时，衍射成象愈宽。在量子力学里，对于粒子的位置与速度的测量精度是互相矛盾的，一个提高了，另一个必降低，即所谓测不准原理。所有这些在谐波分析中反映为宽窄对偶关系，原函数愈窄(即愈"集中")，则谱函数愈宽(即愈"分散")。

表 3.1 中第5项举出了傅氏变换下的尺度伸缩原理：设 $f(t)\sim F(s)$，则有 $f(at)\sim\frac{1}{|a|}F\left(\frac{s}{a}\right)$。当 $a>1$ 时，$f(t)\to f(at)$ 是把 $f(t)$ 的图形横向压缩，而 $F(s)\to\frac{1}{|a|}F\left(\frac{s}{a}\right)$，则是把 $F(s)$ 的图形横向拉伸，伴以纵向压缩而面积不变。因此原函数变窄时，谱函数变宽，即在傅氏变换下"宽度"朝相反方向转化。

至于原函数与自己的谱函数的宽窄对比，则仍可以从最简单的例子

$$\delta(t)\sim 1,\quad \operatorname{rect} t\sim \operatorname{sinc} s$$

看出。δ 函数是集中在一个点的函数，是最窄、最集中的函数，而它的谱函数1则可以说是最宽最分散的函数。$\operatorname{rect} t$ 是矩形，宽度为1。它的谱函数 $\operatorname{sinc} s$ 是波纹状有中峰的函数。虽然直接地无所谓宽度，但可以适当地赋以等效的宽度，例如定义为其面积与峰点的高度之比。由于 $\operatorname{sinc} s$ 的面积为1，峰点即原点处值为1，故等效宽度为1。作尺度变换，$\operatorname{rect}\to\operatorname{rect} at$ (不妨设 $a>1$)，宽度缩至 $\frac{1}{a}$，而相应地谱函数 $\operatorname{sinc} s\to\frac{1}{a}\operatorname{sinc}\frac{s}{a}$，后者面积仍为1而峰点高度为 $\frac{1}{a}$，故等效宽度放大为 a，即新的原函数与谱函数的宽度互为倒数。

上面只是对特定的函数规定了宽度的概念，还太局限，也不尽合理。可以在更广的范围引进比较合理的宽度概念，并使得宽窄对偶性从数量上表达出来。为此，可以仿照力学上定重心和矩量的方法先按下式定义函数 f 的"重心" t_f:

$$t_f\int_{-\infty}^{\infty}|f|^2dt=\int_{-\infty}^{\infty}t|f|^2dt \tag{3.3.77}$$

再对重心点 t_f 求矩量而定义宽度 Δ_f 如下:

$$(\Delta_f)^2\int_{-\infty}^{\infty}|f|^2dt=\int_{-\infty}^{\infty}(t-t_f)^2|f|^2dt \tag{3.3.78}$$

这样的宽度恒正，并有平移不变性和比例压缩性，即

$$\Delta_f>0$$

$$\Delta_{T_af}=\Delta_f,\quad T_af(t)=f(t-a)$$

$$\Delta_{S_af}=\frac{1}{|a|}\Delta_f,\quad S_af(t)=f(at)$$

符合于直观上的宽度概念。相应地，对于 $f(t)$ 的谱函数 $F(s)$ 也有重心 s_F 和宽度 Δ_F:

$$s_F\int_{-\infty}^{\infty}|F|^2ds=\int_{-\infty}^{\infty}s|F|^2ds \tag{3.3.79}$$

$$(\Delta_F)^2\int_{-\infty}^{\infty}|F|^2ds=\int_{-\infty}^{\infty}(s-s_F)^2|F|^2ds \tag{3.3.80}$$

可以建立一个具有普遍意义的不等式

$$f(t)\sim F(s):\ \int_{-\infty}^{\infty}t^2|f|^2dt\cdot\int_{-\infty}^{\infty}s^2|F|^2ds\geqslant\left(\frac{1}{4\pi}\right)^2\int_{-\infty}^{\infty}|f|^2dt\cdot\int_{-\infty}^{\infty}|F|^2ds \quad (3.3.81)$$

并在这个基础上得到关于宽窄对偶性的不等式

$$\Delta_f\cdot\Delta_F\geqslant\frac{1}{4\pi} \quad (3.3.82)$$

量子力学中的测不准原理就是表为这种数学形式，事实上它在谐波分析中是普遍成立的。

§3.4 离散傅氏变换及其快速算法

3.4.1 离散傅氏变换

在有限多个等距离散数据的基础上进行谐波分析的工具是有限傅氏级数或称离散傅氏变换，这就是

正变换

$$U_k=\sum_{j=0}^{N-1}u_jW^{-jk},\ k=0,\ 1,\ \cdots,\ N-1 \quad (3.4.1)$$

逆变换

$$u_j=\frac{1}{N}\sum_{k=0}^{N-1}U_kW^{jk},\ j=0,\ 1,\ \cdots,\ N-1 \quad (3.4.2)$$

它表示了向量$(u_0,\ u_1,\cdots,u_{N-1})$与$(U_0,\ U_1,\ \cdots,U_{N-1})$之间的线性互逆关系。这里采用了记号

$$W=W_N=e^{2\pi i/N} \quad (3.4.3)$$

它是1的一个N次原根，即

$$W^N=1,\quad 并且\ W^j=1,\ 当且仅当\ j\equiv0\ (\mathrm{mod}\ N)^{\ominus} \quad (3.4.4)$$

这个简单性质是离散傅氏变换的基本点。由于

$$0=1-W^{Nk}=(1-W^k)(1+W^k+W^{2k}+\cdots+W^{(N-1)k}),$$

因此有正交关系

$$1+W^k+W^{2k}+\cdots+W^{(N-1)k}=\sum_{j=0}^{N-1}W^{jk}=\begin{cases}N, & k\equiv0\ (\mathrm{mod}\ N)\\ 0, & k\not\equiv0\ (\mathrm{mod}\ N)\end{cases} \quad (3.4.5)$$

据此可以导出(3.4.1)与(3.4.2)的等价性。事实上，设(3.4.2)成立，把它两端各乘以$W^{-jk'}$，并对$j=0,\ 1,\ \cdots,N-1$求和，利用(3.4.5)，

$$\sum_{j=0}^{N-1}u_jW^{-jk'}=\frac{1}{N}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}U_kW^{j(k-k')}=\frac{1}{N}\cdot NU_{k'}=U_{k'}$$

即(3.4.1)成立。也就是说(3.4.1)与(3.4.2)是互逆的公式。

公式(3.4.1)，(3.4.2)不仅对于$k,\ j=0,\ 1,\ \cdots,\ N-1$有意义，而且对$k,\ j$为一切整数也有意义，为此只需理解

$$\begin{aligned}u_j&=u_{j'} \quad 当\ j\equiv j' \quad (\mathrm{mod}\ N)\\ U_k&=U_{k'} \quad 当\ k=k' \quad (\mathrm{mod}\ N)\end{aligned} \quad (3.4.6)$$

两个周期为N的无穷序列

$$u_j,\ U_k,\ j,\ k=0,\ \pm1,\ \cdots$$

事实上都只有N个自由度，即由任意N个相连分量$(u_p,\ u_{p+1},\ \cdots,u_{p+N-1})$，$(U_q,\ U_{q+1},\ \cdots,U_{q+N-1})$，用周期性延拓可以分别决定全序列$\{u_j\}$，$\{U_k\}$，利用式(3.4.6)、(3.4.4)不难验证互逆关系(3.4.1)、(3.4.2)等价于下列互逆关系：

⊖ $j\equiv j'\ (\mathrm{mod}\ N)$的意思是$j-j'$为$N$的整数倍，因此$j\equiv0\ (\mathrm{mod}\ N)$的意思是$j$为$N$的整数倍。

$$U_k=\sum_{j=p}^{p+N-1}u_jW^{-jk},\quad k=q,\ q+1,\ \cdots,\ q+N-1 \tag{3.4.7}$$

$$u_j=\frac{1}{N}\sum_{k=q}^{q+N-1}U_kW^{jk},\quad j=p,\ p+1,\ \cdots,\ p+N-1 \tag{3.4.8}$$

这是 N 点离散傅氏变换的一般形式，在实践中的问题通常提成这样的一般形式；而在实际计算时则利用周期性归化为 $(u_0,\ u_1,\ \cdots,\ u_{N-1})$，$(U_0,\ U_1,\ \cdots,\ U_{N-1})$，并按标准形式(3.4.1)或(3.4.2)执行。

因此所谓离散变换也可以指(3.4.7)及(3.4.8)，而 $u_j,\ U_k,\ j,\ k=0,\ \pm1,\ \cdots$ 恒理解成周期为 N 的无穷序列。

设 $\{u_j\}$ 为实数列，即 $u_j\equiv u_j^*$，则其变换 $\{U_k\}$ 为共轭 $U_{N-k}\equiv U_k^*$，由于周期性 (3.4.3) 这也等价于 $U_{-k}\equiv U_k^*$。事实上

$$U_{N-k}=\sum_{j=0}^{N-1}u_jW_N^{-j(N-k)}=\sum_{j=0}^{N-1}u_jW_N^{jk}=\sum_{j=0}^{N-1}u_j^*(W_N^{-jk})^*=U_k^*$$

类似地，可以证明：当 $\{u_j\}$ 为实数并且对称 $u_{N-j}\equiv u_j$（等价于 $u_{-j}\equiv u_j$）时，$\{U_k\}$ 也是实数并且对称；当 $\{u_j\}$ 是实数并且反对称 $u_{N-j}\equiv -u_j$（等价于 $u_{-j}\equiv -u_j$）时，则 $\{U_k\}$ 是纯虚数、反对称。归纳起来，上述结论可以表为

$$u_j\equiv u_j^*\sim U_k\equiv U_{N-k}^* \tag{3.4.9}$$

$$u_j\equiv u_j^*\equiv u_{N-j}\sim U_k\equiv U_k^*\equiv U_{N-k} \tag{3.4.10}$$

$$u_j\equiv u_j^*\equiv -u_{N-j}\sim U_k\equiv -U_k^*\equiv -U_{N-k} \tag{3.4.11}$$

离散傅氏变换(3.4.1)、(3.4.2)可以写成

$$U_k=\sum_{j=0}^{N-1}u_j\cos(2\pi jk/N)-i\sum_{j=0}^{N-1}u_j\sin(2\pi jk/N)$$

$$u_j=\frac{1}{N}\sum_{k=0}^{N-1}U_k\cos(2\pi jk/N)+i\frac{1}{N}\sum_{k=0}^{N-1}U_k\sin(2\pi jk/N)$$

不妨定义离散的余弦及正弦变换为

$$\begin{cases}u_k^c=\sum_{j=0}^{N-1}u_j\cos(2\pi jk/N),\ k=0,\ \cdots,\ N-1\\ u_k^s=\sum_{j=0}^{N-1}u_j\sin(2\pi jk/N),\ k=0,\ \cdots,\ N-1\end{cases} \tag{3.4.12}$$

于是

$$U_k=u_k^c-iu_k^s,\quad k=0,\ \cdots,\ N-1$$

$$u_j=\frac{1}{N}U_j^c+i\frac{1}{N}U_j^s=\frac{1}{N}(u_j^{cc}+u_j^{ss})+i\frac{1}{N}(-u_j^{sc}+u_j^{cs}),\quad j=0,\ \cdots,\ N-1$$

当 $\{u_j\}$ 为实数时，上式简化为

$$U_k=u_k^c-iu_k^s,\quad k=0,\ \cdots,\ N-1$$

$$u_j=\frac{1}{N}(u_j^{cc}+u_j^{ss}),\quad j=0,\ \cdots,\ N-1$$

当 $\{u_j\}$ 为实数、对称时，$\{U_k\}$ 也是实数、对称(3.4.10)。傅氏变换可以单独用余弦变换表达

$$\begin{cases}U_k=u_k^c,\quad k=0,\ \cdots,\ N-1\\ u_j=\frac{1}{N}U_j^c=\frac{1}{N}u_j^{cc},\quad j=0,\ \cdots,\ N-1\end{cases} \tag{3.4.13}$$

由于对称性只需算一半的分量:

当 $N=2N'$ 时

$$u_k^c=\sum_{j=0}^{N'} a_j u_j \cos(2\pi jk/N), \quad k=0, \cdots, N'$$

$$a_0=1, \quad a_1=\cdots=a_{N'-1}=2, \quad a_{N'}=1$$

当 $N=2N'+1$ 时

$$u_k^c=\sum_{j=0}^{N'-1} a_j u_j \cos(2\pi jk/N), \quad k=0, \cdots, N'-1$$

$$a_0=1, \quad a_1=\cdots=a_{N'-1}=2$$

当 $\{u_j\}$ 为实数、反对称时，$\{U_k\}$ 是纯虚、反对称式(3.4.11)。傅氏变换可以单独用正弦变换来表达

$$\begin{aligned} iU_k&=u_k^s, k=0, \cdots, N-1 \\ u_j&=\frac{1}{N} u_j^{ss}, \quad j=0, \cdots, N-1 \end{aligned} \tag{3.4.14}$$

这时由于反对称性，也只需计算一半的分量：无论对于 $N=2N'$ 或 $2N'+1$ 都有

$$u_k^s=2\sum_{j=1}^{N'-1} u_j \sin(2\pi jk/N), \quad k=1, \cdots, N'-1$$

3.4.2 离散卷积

对于两个序列 u_j, $v_j(j=0, 1, \cdots, N-1)$ 定义其离散卷积，即新的序列

$$w_j=\sum_{j'=0}^{N-1} u_{j'} v_{j-j'}, \quad j=0, 1, \cdots, N-1 \tag{3.4.15}$$

这里 u_j, v_j 都理解为对下标按周期 N 作拓展，因此有

$$\begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_{N-1} \end{pmatrix}=\begin{pmatrix} v_0 & v_{N-1} & v_{N-2} & \cdots & v_1 \\ v_1 & v_0 & v_{N-1} & \cdots & v_2 \\ v_2 & v_1 & v_0 & \cdots & v_3 \\ \cdots & \cdots & \cdots & & \cdots \\ v_{N-1} & v_{N-2} & v_{N-3} & \cdots & v_0 \end{pmatrix}\begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{pmatrix} \tag{3.4.16}$$

而 w_j 也同样具有周期 N。

图3.12表示这种卷积的形式，其中(a), (b)表示原函数 u, v 及其周期延拓，(c)表示 v 的反转，然后依次右移 0, 1, $\cdots$, $N-1$ 格，并分别乘、加，即得(f)中的卷积值 w。

很容易验证，与连续的情况相类似，这种周期性离散卷积与"逐点"乘积在离散傅氏变换下是互换的。设

$$u_j \sim U_k, \quad v_j \sim V_k \tag{3.4.17}$$

则有

$$\sum_{j'=0}^{N-1} u_{j'} v_{j-j'} \sim U_k V_k \tag{3.4.18}$$

$$u_j v_j \sim \frac{1}{N} \sum_{k'=0}^{N-1} U_{k'} V_{k-k'} \tag{3.4.19}$$

为了方便，离散傅氏变换的一些主要性质列在表 3.3 备考。所有这些性质都是初等的，不难根据离散傅氏变换的定义和三角函数的周期性加以验证。

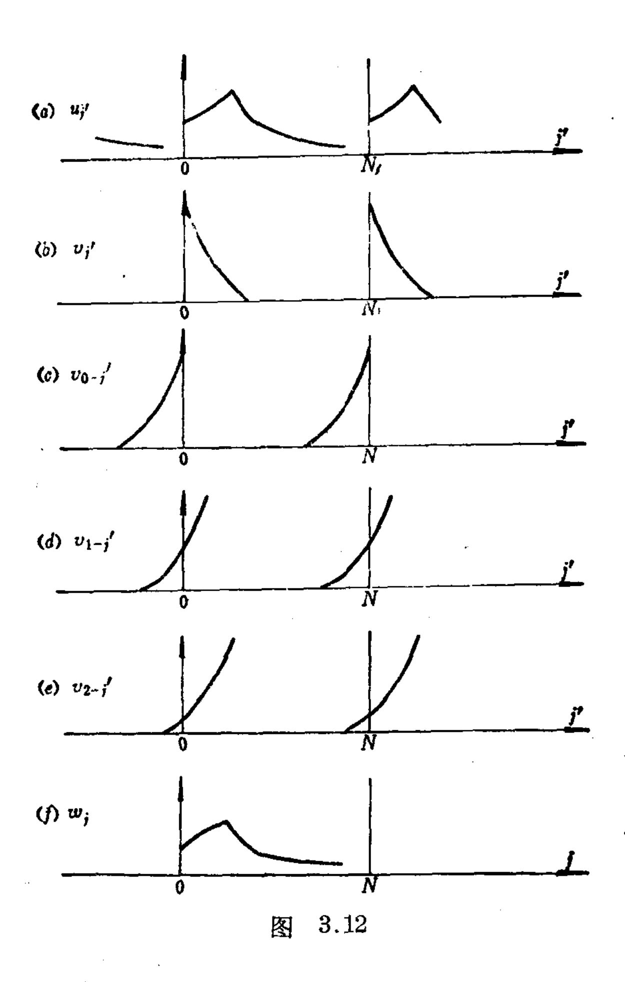


图 3.12

表3.3 离散傅氏变换性质简表

	原 函 数	谱 函 数	说 明
	u_j v_j	U_k V_k	$j, k=0, 1, \cdots N-1$
1	U_j	u_{N-k}	互逆性
2	$\alpha u_j+\beta v_j$	$\alpha U_k+\beta V_k$	线性迭加原理
3	u_{N-j}	U_{N-k}	偶(奇)函数～偶(奇)函数
4	u_j^*	U_{N-k}^*	实函数 $f_j=f_j^*$～$F_k=F_{N-k}^*$
5	$u_{j-j'}$	$W_N^{j'k}U_k$	平移距离 j'(整数)～乘因子 $W_N^{j'k}$
6	$\sum_{j'=0}^{N-1} u_{j'}v_{j-j'}$ u_jv_j	$U_k \cdot V_k$ $\frac{1}{N}\sum_{k'=0}^{N-1} U_{k'}V_{k-k'}$	卷积～乘积
7	$\sum_{j=1}^{N-1} u_j^* v_j=\frac{1}{N}\sum_{k'=0}^{N-1} U_k^* V_k$ $\sum_{j=0}^{N-1} \|u_j\|^2=\frac{1}{N}\sum_{k=0}^{N-1} \|U_k\|^2$		内积不变性

3.4.3 快速傅氏变换

离散傅氏变换的计算，就是从已知的$(u_0, u_1, \cdots, u_{N-1})$按显式(3.4.1)计算未知的$(U_0, U_1, \cdots, U_{N-1})$，或者从已知的$(U_0, U_1, \cdots, U_{N-1})$按显式(3.4.2)计算未知的$(u_0, u_1, \cdots, u_{N-1})$。表面上看，这是一个极其简单的"算术"问题。直接根据(3.4.1)或(3.4.2)可以看到，计算一个 N 点离散傅氏变换的工作量为 N^2 个复运算㊀——所谓一个复运算是指一个复数乘法连同一个复数加法。对这种直接算法，考虑到三角函数的对称性等等因素，可以把工作量节约一定的倍数，但量级仍旧为 N^2 这样的工作量。对于实际谐波分析中较大的 N，特别对于"实时"的计算，这往往成为沉重的负担。正因如此，在相当长的时间内，在各种领域中的谐波分析的问题中数值手段没有得到广泛应用，人们往往更多地采用物理模拟的手段。近年来在生产实践的基础上逐渐明确并发展了逐次分半或类似的递推算法，使得计算一个 N 点变换的工作量降至 $N\log_2 N$ 个复运算，比传统的直接算法提高工效 $N/\log_2 N$ 倍。当 N 很大时(实践上往往如此)这是数量级上的提高。例如 $N=16$ 时为 4 倍，$N\approx 10^3$ 时约为 100 倍，$N\approx 10^6$(有这样的情况)时约为五万倍！这样，基本上克服了所谓"时间域"与"频率域"转换中的计算障碍，从而为数值谐波分析方法在科学技术的许多方面(例如光谱和声谱分析、全息技术、地震勘探、数字信号处理、图形信息处理、微分方程数值解等等)的广泛应用开辟了道路。通常把这类工作量为 $N\log_2 N$ 的算法统称为快速傅氏变换(FFT)。这类方法除了有快速的特点外，还有下列特点：精确度比传统方法所得到的高、计算过程稳定、比较简单等。目前它在数值谐波分析中已占主导地位。下面将介绍其中主要的一种，即逐次分半算法。

在逐次分半法中恒取 $N=2^m$，并把一次 N 点变换的过程分解为 $m=\log_2 N$ 步，每步计算一个简化的傅氏变换，工作量为 N 个复运算，因而总工作量为 $N\log_2 N$。由于分点数 N 的选取在多数场合下人们是有主动权的，因此取 $N=2^m$ 这一特殊形式并不是严重的限制。现对于

$$F_k=\sum_{j=0}^{N-1} f_j W^{-jk}, \quad k=0, 1, \cdots, N-1 \tag{3.4.20}$$

以 $N=2^3=8$ 为例说明算法思想。

将下标 $j, k=0, 1, \cdots, 7$ 表为二进制

$$j=(j_2, j_1, j_0)=j_2\cdot 2^2+j_1\cdot 2+j_0, \quad j_0, j_1, j_2=0, 1$$
$$k=(k_2, k_1, k_0)=k_2\cdot 2^2+k_1\cdot 2+k_0, \quad k_0, k_1, k_2=0, 1$$

由于

$$\begin{aligned} j\cdot k &=(j_2, j_1, j_0)\cdot(k_2, k_1, k_0)=(j_2\cdot 2^2+j_1\cdot 2+j_0)(k_2\cdot 2^2+k_1\cdot 2+k_0) \\ &=j_2(k_2\cdot 2^4+k_1\cdot 2^3+k_0\cdot 2^2)+j_1(k_2\cdot 2^3+k_1\cdot 2^2+k_0\cdot 2)+j_0(k_2\cdot 2^2+k_1\cdot 2+k_0) \\ &\equiv j_2(k_0\cdot 2^2)+j_1(k_1\cdot 2^2+k_0\cdot 2^1)+j_0(k_2\cdot 2^2+k_1\cdot 2+k_0) \pmod{2^3} \end{aligned}$$

以及(3.4.3)故有

$$W^{-jk}=W^{-j_2(k_0, 0, 0)}W^{-j_1(k_1, k_0, 0)}W^{-j_0(k_2, k_1, k_0)}$$

命 $F_k=F(k_2, k_1, k_0)$，$f_j=f(j_2, j_1, j_0)$，于是

㊀ 这里不计及三角函数 $W^{jk}=\cos 2\pi jk/N+i\sin 2\pi jk/N$ 的产生。

$$F(k_2, k_1, k_0)=\sum_{j=0}^{2^3-1}f_jW^{-jk}=\sum_{j_0=0}^{1}\sum_{j_1=0}^{1}\sum_{j_2=0}^{1}f(j_2, j_1, j_0)W^{-j_2(k_0,0,0)}W^{-j_1(k_1,k_0,0)}W^{-j_0(k_2,k_1,k_0)}$$

$$=\sum_{j_0=0}^{1}\left(\sum_{j_1=0}^{1}\left(\sum_{j_2=0}^{1}f(j_2, j_1, j_0)W^{-j_2(k_0,0,0)}\right)W^{-j_1(k_1,k_0,0)}\right)W^{-j_0(k_2,k_1,k_0)}$$

这样就导出了递推过程

$$f^{(0)}(j_2, j_1, j_0)=f(j_2, j_1, j_0),\quad j_2, j_1, j_0=0, 1$$

$$f^{(1)}(k_0, j_1, j_0)=\sum_{j_2=0}^{1}f^{(0)}(j_2, j_1, j_0)W^{-j_2(k_0,0,0)}$$

$$=f^{(0)}(0, j_1, j_0)+f^{(0)}(1, j_1, j_0)W^{-(k_0,0,0)},\quad k_0, j_1, j_0=0, 1$$

$$f^{(2)}(k_0, k_1, j_0)=\sum_{j_1=0}^{1}f^{(1)}(k_0, j_1, j_0)W^{-j_1(k_1,k_0,0)}$$

$$=f^{(1)}(k_0, 0, j_0)+f^{(1)}(k_0, 1, j_0)W^{-(k_1,k_0,0)},\quad k_0, k_1, j_0=0, 1$$

$$f^{(3)}(k_0, k_1, k_2)=\sum_{j_0=0}^{1}f^{(2)}(k_0, k_1, j_0)W^{-j_0(k_2,k_1,k_0)}$$

$$=f^{(2)}(k_0, k_1, 0)+f^{(2)}(k_0, k_1, 1)W^{-(k_2,k_1,k_0)},\quad k_0, k_1, k_2=0, 1$$

$$F(k_2, k_1, k_0)=f^{(3)}(k_0, k_1, k_2),\quad k_2, k_1, k_0=0, 1$$

把下标改写即得

开　始：$f^{(0)}(k_2, k_1, k_0)=f(k_2, k_1, k_0)$

第一步：$f^{(1)}(k_2, k_1, k_0)=f^{(0)}(0, k_1, k_0)+f^{(0)}(1, k_1, k_0)W^{-(k_2,0,0)}$

第二步：$f^{(2)}(k_2, k_1, k_0)=f^{(1)}(k_2, 0, k_0)+f^{(1)}(k_2, 1, k_0)W^{-(k_1,k_2,0)}$

第三步：$f^{(3)}(k_2, k_1, k_0)=f^{(2)}(k_2, k_1, 0)+f^{(2)}(k_2, k_1, 1)W^{-(k_0,k_1,k_2)}$

结　束：$F(k_2, k_1, k_0)=f^{(3)}(k_0, k_1, k_2)$

$$k_0, k_1, k_2=0, 1$$

注意这个算法实现的关键在于下标的组织，在结束步要作下标的按位反转$(k_2, k_1, k_0)\rightarrow(k_0, k_1, k_2)$，即

$$k=(k_2, k_1, k_0)\rightarrow(k_0, k_1, k_2)=k'$$
$$0=(0, 0, 0)\rightarrow(0, 0, 0)=0$$
$$1=(0, 0, 1)\rightarrow(1, 0, 0)=4$$
$$2=(0, 1, 0)\rightarrow(0, 1, 0)=2$$
$$3=(0, 1, 1)\rightarrow(1, 1, 0)=6$$
$$4=(1, 0, 0)\rightarrow(0, 0, 1)=1$$
$$5=(1, 0, 1)\rightarrow(1, 0, 1)=5$$
$$6=(1, 1, 0)\rightarrow(0, 1, 1)=3$$
$$7=(1, 1, 1)\rightarrow(1, 1, 1)=7$$

此外每一步 W^{-1} 的幂次码$(k_2, 0, 0)$，$(k_1, k_2, 0)$，(k_0, k_1, k_2)也需要由(k_2, k_1, k_0)作按位反转再分别左移 2 位、1 位及 0 位得到

$$(k_2, k_1, k_0)\xrightarrow{\text{反转}}(k_0, k_1, k_2)\begin{cases}\xrightarrow{\text{左移 2 位}}(k_2, 0, 0)\\ \xrightarrow{\text{左移 1 位}}(k_1, k_2, 0)\\ \xrightarrow{\text{左移 0 位}}(k_0, k_1, k_2)\end{cases}$$

因此一个方便的办法是在开始步作反转，则以后就无须再作反转，即算法改为

$$f^{(0)}(k_2,\ k_1,\ k_0)=f(k_0,\ k_1,\ k_2)$$
$$f^{(1)}(k_2,\ k_1,\ k_0)=f^{(0)}(k_2,\ k_1,\ 0)+f^{(0)}(k_2,\ k_1,\ 1)W^{-(k_0,0,0)}$$
$$f^{(2)}(k_2,\ k_1,\ k_0)=f^{(1)}(k_2,\ 0,\ k_0)+f^{(1)}(k_2,\ 1,\ k_0)W^{-(k_1,k_0,0)}$$
$$f^{(3)}(k_2,\ k_1,\ k_0)=f^{(2)}(0,\ k_1,\ k_0)+f^{(2)}(1,\ k_1,\ k_0)W^{-(k_2,k_1,k_0)}$$
$$F(k_2,\ k_1,\ k_0)=f^{(3)}(k_2,\ k_1,\ k_0)$$
$$k_0,\ k_1,\ k_2=0,\ 1 \tag{3.4.21}$$

这里每一级的 W^{-1} 的幂次 $(k_0,\ 0,\ 0)$, $(k_1,\ k_0,\ 0)$, $(k_2,\ k_1,\ k_0)$ 直接由 $(k_2,\ k_1,\ k_0)$ 左移 2、1、0 位而得。图 3.13 给出这个三级计算各分量的流程图，其中虚线表示直接相加，实线表示乘以 W^{-1} 的相当幂次后相加。在每一级的计算中都是把变换的分量两两配对，每一对

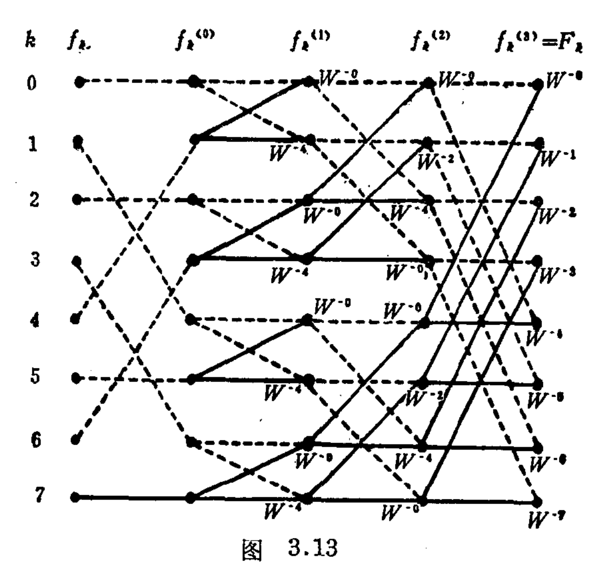


图 3.13

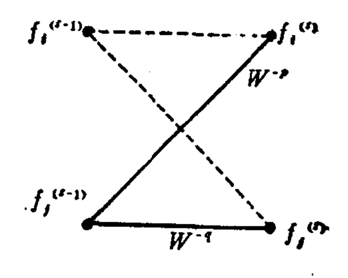


图 3.14

是一个"蝴蝶结"形的计算(图 3.14)：

$$f_i^{(s-1)}+f_j^{(s-1)}\cdot W^{-p}=f_i^{(s)}$$
$$f_i^{(s-1)}+f_j^{(s-1)}\cdot W^{-q}=f_j^{(s)}$$

这一算法显然可以推广到一般的 $N=2^m$。从 0 到 $N-1$ 的任意整数 k 可以表为二进制

$$k=(k_{m-1},\ \cdots,\ k_1,\ k_0)=k_{m-1}\cdot 2^{m-1}+\cdots+k_1\cdot 2+k_0\cdot 1$$
$$k_{m-1},\ \cdots,\ k_1,\ k_0=0,\ 1;\ \ 0\leqslant k\leqslant N-1$$

于是有

$$f^{(0)}(k_{m-1},\ \cdots,\ k_1,\ k_0)=f(k_0,\ k_1,\ \cdots,\ k_{m-1})$$
$$f^{(s)}(k_{m-1},\ \cdots,\ k_1,\ k_0)=f^{(s-1)}(k_{m-1},\ \cdots,\ k_s,\ 0,\ k_{s-2},\ \cdots,\ k_0)$$
$$+f^{(s-1)}(k_{m-1},\ \cdots,\ k_s,\ 1,\ k_{s-2},\ \cdots,\ k_0)\cdot W^{-(k_{s-1},\cdots,k_0,0,\cdots,0)},\ s=1,\ 2,\ \cdots,\ m$$
$$F(k_{m-1},\ \cdots,k_1,\ k_0)=f^{(r)}(k_{m-1},\ \cdots,\ k_1,\ k_0),\ k_{m-1},\ \cdots,\ k_1,\ k_0=0,\ 1 \tag{3.4.22}$$

这里，每步每个分量需要 1 个复乘和一个复加，即一个所谓"复运算"，N 个分量则要 N 个复运算。共有 $m=\log_2 N$ 步，总计为 $N\cdot m=N\ln_2 N$ 个复运算。这是一个简单而高效的算法。实践表明，即使对于很大的 N，计算过程也是稳定的。

实函数的傅氏变换算法

以上快速算法是对于复变量的。在实际应用中，多数情况下原函数是实数(它的傅氏变换一般还是复的)，当然可以直接套用上列标准算法，但不很经济。设有实数列 $u_j\ (j=0,$

$1, \cdots, N-1$)。由于

$$u_j = u_j^*, \quad j=0, 1, \cdots, N-1$$

其傅氏变换 U_k 必满足关系式

$$U_k = U_{N-k}^*, \quad k=0, 1, \cdots, N-1 \tag{3.4.23}$$

因此有节约的潜力。下面举的两种处理方法，都是间接套用标准算法，但工作量可以节约一半。

1）用一个 N 点复变换来计算两个 N 点实变换 $a_j \sim A_k$, $b_j \sim B_k$。

作复向量

$$c_j = a_j + ib_j, \quad j=0, 1, \cdots, N-1$$

套用标准算法 $c_j \sim C_k$, $k=0, 1, \cdots, N-1$。由线性迭加原理得 $C_k = A_k + iB_k$。

另一方面，因为 $A_k = A_{N-k}^*$, $B_k = B_{N-k}^*$, 因此有

$$C^*_{N-k} = A_k - iB_k$$

于是

$$\left.\begin{aligned} A_k &= \frac{1}{2}(C_k + C_{N-k}^*) \\ B_k &= \frac{1}{2i}(C_k - C_{N-k}^*) \end{aligned}\right\} \quad k=0, 1, \cdots, N-1 \tag{3.4.24}$$

2）用一个 N 点复变换来计算一个 $2N$ 点实变换 $u_j \sim U_k$。

作复向量 $$c_j = a_j + ib_j$$

式中 $$a_j = u_{2j}, \quad b_j = u_{2j+1}, \quad j=0, 1, \cdots, N-1$$

套用标准算法 $$c_j \sim C_k = A_k + iB_k, \quad j, k=0, 1, \cdots, N-1$$

此处 A_k, B_k 用公式(3.4.24)计算。又由于

$$U_k = \sum_{j=0}^{2N-1} u_j W_{2N}^{-jk} = \sum_{j=0}^{N-1} (u_{2j} W_{2N}^{-2jk} + u_{2j+1} W_{2N}^{-(2j+1)k})$$

$$= \sum_{j=0}^{N-1} (a_j W_N^{-jk} + b_j W_N^{-jk} \cdot W_{2N}^{-k}) = A_k + B_k W_{2N}^{-k}, \quad k=0, 1, \cdots, 2N-1$$

由于

$$A_{N+k} = A_k, \quad B_{N+k} = B_k$$

$$W_{2N}^{-(N+k)} = W_{2N}^{-k} \cdot W_{2N}^{-N} = -W_{2N}^{-k}$$

因此有

$$\left.\begin{aligned} U_k &= A_k + B_k W_{2N}^{-k} \\ U_{N+k} &= A_k - B_k W_{2N}^{-k} \end{aligned}\right\} \quad k=0, 1, \cdots, N-1 \tag{3.4.25}$$

关于快速算法的其它形式以及对于高维的推广等等可以参考[2]、[3]。

计算 N 点复数列变换以及用它来计算 $2N$ 点实数列变换的语言程序列在本节之末。

利用离散卷积与乘积的互换定理 (3.4.17), (3.4.18), (3.4.19) 可以得到卷积的快速算法。事实上，如果按照公式(3.4.15)"直接"计算 N 点卷积的运算量为 N^2, 但是，如果采用"间接"的方法，即先算 u, v 的傅氏变换，再逐点相乘，然后作逆傅氏变换同样也得卷积而运算量是 $3N\log_2 N$。

在实际的卷积计算中，往往要求 u, v 在 $j=0, 1, \cdots, N-1$ 以外恒为 0, 而不是周期性的。对于非周期卷积，离散的互换定理是不成立的。设 v_j 只当 $j=0, \cdots, L$ 时不为 0, 为了套用傅氏变换方法可以把原程序 u_j, v_j ($j=0, \cdots, N-1$) 的尾部增补 L 个 0 而把周期 N

延成 $N'=N+L$, 则所得周期为 N' 的卷积的前 N 个分量，就是非周期性卷积的准确值。关于卷积计算的细节可以参考[3]。

快速傅氏变换程序

逐次分半算法的程序实现可以有多种多样。下面介绍一个计算 $2N$ 点实数列和 N 点复数列的快速算法语言程序，这个程序比较简短，比较节省内存，但不是最快的。

说 明

1 过程 FFRE(A, B, M) 用来计算实序列 $x_0, x_1, \cdots, x_{2N-1}(N=2^M)$ 的余弦及正弦变换

$$\begin{aligned} a_k&=\sum_{j=0}^{2N-1} x_j\cos(2\pi jk/N) \\ b_k&=\sum_{j=0}^{2N-1} x_j\sin(2\pi jk/N) \end{aligned} \qquad k=0, 1, \cdots, 2N-1$$

而 $x_0, x_1, \cdots, x_{2N-1}$ 的离散(复)傅氏变换就是

$$X_k=a_k-ib_k,\quad k=0, 1, \cdots, 2N-1$$

在计算时首先从实数列 $x_0, \cdots, x_{2N-1}$ 作成复数列 $z_0, \cdots, z_{N-1}$:

$$z_j=x_{2j}+ix_{2j+1},\quad j=0, 1, \cdots, N-1$$

将 $\{z_j\}$ 的实部 $\{x_{2j}\}$ 放在场 $A[j]$ 中，虚部 $\{x_{2j+1}\}$ 放在场 $B[j]$ 中，调用 FFRE(A, B, M)后结果在 A, B 中，分别是 $\{2a_k\}$, $\{-2b_k\}$, $k=0, 1, \cdots, N-1$。

过程 FFRE 是利用计算 N 点复数列傅氏变换的快速方法来计算 $2N$ 点的实数列的傅氏变换。

2 过程 FFT(A, B, M) 也可单独使用以计算复数列 $\{y_j\}$, $j=0, 1, \cdots, N-1(N=2^M)$ 的(复)傅氏变换

$$Y_k=\sum_{j=0}^{N-1} y_j e^{-2\pi ijk/N},\quad k=0, 1, \cdots, N-1$$

为此先要把复数列 $\{y_j\}$ 的实数列 $\{\mathrm{Re}\,y_j\}$ 和虚数列 $\{\mathrm{Im}\,y_j\}$ 按下标的自然顺序分别放在场 A 与 B 中。计算结果在 A 中为 $\{Y_k\}$ 的实数列 $\{\mathrm{Re}Y_k\}$, B 中为虚数列 $\{\mathrm{Im}Y_k\}$, $k=0, 1, \cdots, N-1$。

在过程 FFT 中，语句 L1 至 L7 实现"就地"的按位反转，其后实现"就地"的快速算法，在每个迭代步计算一次所需的三角函数，不另占单元。

```
过程  FFRE(A, B, M); 场 A, B; 简变 M, N;
始  过程 FFT(A, B, M); 场 A, B; 简变 M, N;
    始  简变 K1, U1, U2, W1, W2, T1, T2, NV2, NM1, J, K, IP, LE1, LE, PI;
        0⇒J; 2↑M⇒N; N/2⇒NV2; N-1⇒NM1;
        对于 I=0 到 NM1-1 步长 1 执行
        始  若 J≤I 则转 L5 否; A[J]⇒T1; B[J]⇒T2; A[I]⇒A[J]; B[I]⇒B[J];
            T1⇒A[I]; T2⇒B[I];
        L5: NV2⇒K;
        L6: 若 J<K 则转 L7 否; J-K⇒J; K/2⇒K; 转 L6;
        L7: J+K⇒J;
        终;
```

```
    3.1415926⇒PI;
    对于 L=1 到 M 步长 1 执行
    始  2↑L⇒LE; LE/2⇒LE1; 1⇒U1; 0⇒U2; §COS(PI/LE1)⇒W1;
        §SIN(PI/LE1)⇒W2;
        对于 J=0 到 LE1-1 步长 1 执行
        始  对于 I=J 到 N-1 步长 LE 执行
            始  I+LE1⇒IP; A[IP]*U1+B[IP]*U2⇒T1;
                B[IP]*U1-A[IP]*U2⇒T2; A[I]-T1⇒A[IP];
                B[I]-T2⇒B[IP]; A[I]+T1⇒A[I]; B[I]+T2⇒B[I];
            终;
            U1*W1-U2*W2⇒K1; U1*W2+U2*W1⇒U2; K1⇒U1;
        终;
    终;
终;
过程 REAL(A, B, N); 场 A, B; 简变 N;
始  简变 AA, AB, BA, BB, RE, IM, CD, NH, CN, SN, SD, RAD, R, K;
    §ENTI(N/2)⇒NH; 3.1415926/N⇒RAD; §SIN(RAD)⇒SD;
    -(2*§SIN (0.5*RAD))↑2⇒R; -0.5*R⇒CD; 1⇒CN; 0⇒SN; -SD⇒SD;
    A[0]⇒A[N]; B[0]⇒B[N];
    对于 J=0 到 NH 步长 1 执行
    始 N-J⇒K;
      A[J]+A[K]⇒AA; A[J]-A[K]⇒AB; B[J]+B[K]⇒BA;
      B[J]-B[K]⇒BB; CN*BA+SN*AB⇒RE; SN*BA-CN*AB⇒IM;
      IM-BB⇒B[K]; IM+BB⇒B[J]; AA-RE⇒A[K]; AA+RE⇒A[J];
      R*CN+CD⇒CD; CD+CN⇒CN; R*SN+SD⇒SD; SD+SN⇒SN;
    终;
终;
FFT(A, B, M); REAL(A, B, N);
终
```

§3.5 取样效应

当连续的谐波分析代以离散化的谐波分析时，通常不可避免引起误差。这是因为，第一，连续变量被代以离散取样；第二，变量的无穷范围被代以有限的范围。前者导致所谓频率混叠效应，后者导致所谓谱线渗漏效应。这两项构成了离散化谐波分析的误差的主要来源。下面分别加以初步的分析，这对于傅氏变换的近似计算有指导意义。

3.5.1 离散取样与频谱混叠效应

在等距离散点 $t=j\cdot\Delta t(j=0,\ \pm1,\ \cdots)$ 的基础上来观察连续变量的函数 $f(t)$ 时，由于丢失了信息，一般要导致频谱失真。现在来讨论这种误差。

对于两个不同频率 s, s' 的谐波 $f_s(t)=e^{2\pi ist}$, $f_{s'}(t)=e^{2\pi is't}$ 分别得到离散序列 $f_s(j\Delta t)$, $f_{s'}(j\Delta t)$, $j=0, \pm1, \cdots$。如果频差 $s-s'$ 为取样频率 $\sigma=\frac{1}{\Delta t}$ 的整数倍，即

$$s=s'+\frac{m}{\Delta t},\quad m=\text{整数} \tag{3.5.1}$$

则两组观察值完全重合：

$$f_s(j\Delta t)=e^{2\pi isj\Delta t}=e^{2\pi i\left(s'+\frac{m}{\Delta t}\right)j\Delta t}=e^{2\pi is'j\Delta t}=f_{s'}(j\Delta t),\ j=0,\ \pm1,\ \cdots$$

因此当以 Δt 为间距取样时，这样两个频率 s 与 s' 就完全混同起来，不论如何加工处理都无从分辨。由于任何频率 s 一定可以表为

$$s=s'+\frac{m}{\Delta t},\quad |s'|\leqslant\frac{1}{2\Delta t},\quad m=\text{整数} \tag{3.5.2}$$

因此只能辨认低频段 $|s|\leqslant\frac{1}{2\Delta t}$，而其它高频都按照(3.5.2)被折合到这个低频段上来。这相当于把频率轴 $-\infty<s<\infty$ 卷绕在相切于原点 $s=0$ 的周长为 $\frac{1}{\Delta t}$ 的圆周上，$-\frac{1}{2\Delta t}\leqslant s'\leqslant\frac{1}{2\Delta t}$，如图 3.15。

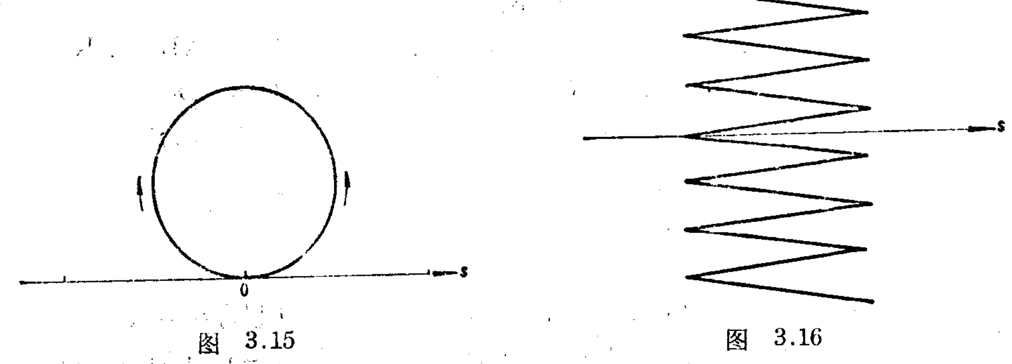


图 3.15　　　　图 3.16

由于 $e^{\pm 2\pi ist}=\cos 2\pi st\pm i\sin 2\pi st$，数学上的正负频率 s，$-s$ 实际上相当于一个正频率 s。因此也可以看作把频率轴反复折叠于 $0\leqslant s\leqslant\frac{1}{2\Delta t}$，如图 3.16。这就是等距离散取样导致的频谱混叠(aliasing)效应。概言之，由于取样间距 Δt 不能是无穷小而只能是有限小，于是频率辨认的范围不能是无穷大而只能是有限大 $|s|\leqslant\frac{1}{2\Delta t}$。频率 $\frac{1}{2\Delta t}$ 通常叫做折叠频率或奈奎斯特频率。

对于一般情况，设 $f(t)\sim F(s)$，即

$$f(t)=\int_{-\infty}^{\infty}F(s)e^{2\pi ist}ds,\quad F(s)=\int_{-\infty}^{\infty}f(t)e^{-2\pi ist}dt \tag{3.5.3}$$

为了方便，命

$$S=\frac{1}{\Delta t} \tag{3.5.4}$$

于是

$$f(j\Delta t)=f(j/S)=\int_{-\infty}^{\infty}F(s)e^{2\pi ijs/S}ds=\sum_{m=-\infty}^{\infty}\int_{\left(m-\frac{1}{2}\right)S}^{\left(m+\frac{1}{2}\right)S}F(s)e^{2\pi ijs/S}ds$$

$$=\sum_{m=-\infty}^{\infty}\int_{-S/2}^{S/2}F(s'-mS)e^{2\pi ij(s'-mS)/S}ds'=\sum_{m=-\infty}^{\infty}\int_{-S/2}^{S/2}F(s-mS)e^{2\pi ijs/S}ds$$

命

$$F_S(s)=\sum_{m=-\infty}^{\infty}F(s-mS),\quad -\infty<s<\infty \tag{3.5.5}$$

这是由 $F(s)$ 逐次平移距离 S 再叠加而得的函数，显然具有周期 S：

$$F_S(s+S)\equiv F_S(s),\quad -\infty<s<\infty \tag{3.5.6}$$

于是有

$$f(j\Delta t)=\int_{-S/2}^{S/2}F_S(s)e^{2\pi ijs/S}ds \tag{3.5.7}$$

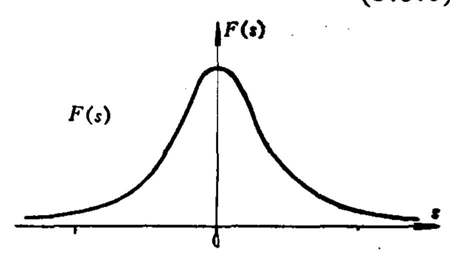
更由于

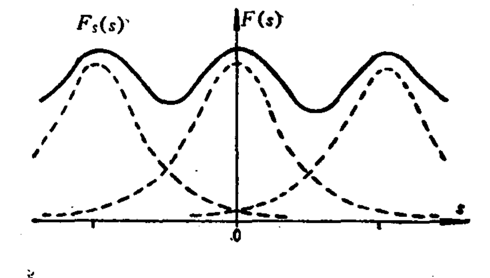
$$\operatorname{rect}s/S=\begin{cases}1, & |s|<S/2\\ 0, & |s|>S/2\end{cases} \tag{3.5.8}$$

因此

$$f(j\Delta t)=\int_{-\infty}^{\infty}(\operatorname{rect}s/S)F_S(s)e^{2\pi isj\Delta t}ds,$$
$$j=0,\ \pm1,\ \pm2,\ \cdots \tag{3.5.9}$$

这就表示，在离散点列 $t=j\Delta t$ 上观察波形 $f(t)$ 时，能够看到的频率范围只是 $|s|\leqslant S/2=1/2\Delta t$，其外的频率都按照(3.5.2)被折合到上述频段，即原来的谱函数被代以折叠了的谱函数 $\operatorname{rect}s/S\cdot F_S(s)$。图 3.17 表示三个函数 $F(s)$、$F_S(s)$ 和 $\operatorname{rect}s/S\cdot F_S(s)$ 之间的关系。可以设想把 s-F 平面卷绕于周长为 S 相切于纵轴 $s=0$ 的圆柱上，于是图线 $F(s)$ 就变为 $\operatorname{rect}s/S\cdot F_S(s)$。





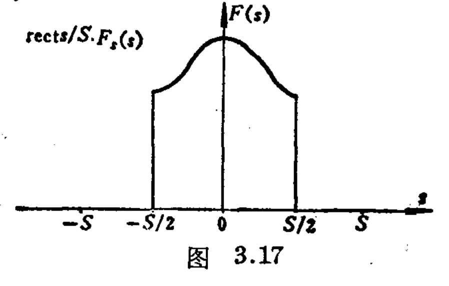


图 3.17

$F_S(s)$ 是周期为 S 的函数，因此可以展为傅氏级数

$$F_S(s)=\sum_{j=-\infty}^{\infty}c_je^{-2\pi ijs/S}$$

$$c_j=\frac{1}{S}\int_{-S/2}^{S/2}F_S(s)e^{2\pi ijs/S}ds$$

由于(3.5.7)

$$c_j=\frac{1}{S}f(j\Delta t)=\Delta t\cdot f(j\Delta t)$$

因此有

$$F_S(s)=\Delta t\sum_{j=-\infty}^{\infty}f(j\Delta t)e^{-2\pi isj\Delta t} \tag{3.5.10}$$

实践上一类重要的情况是 $f(t)$ 的谱函数 $F(s)$ 为紧凑；这时我们称 $f(t)$ 为有限谱宽函数，即 f 的频谱有上限 L 使得

$$F(s)\equiv0\quad 当\quad |s|\geqslant L \tag{3.5.11}$$

这时，只需取样间距 Δt 足够小，即满足条件

$$\Delta t\leqslant1/2L \tag{3.5.12}$$

(也就是说 $S/2=1/2\Delta t\geqslant L$)时，就没有混叠，即

$$F(s)\equiv\operatorname{rect}s/S\cdot F_S(s),\quad -\infty<s<\infty \tag{3.5.13}$$

这从图 3.18 也可以看出，并可与有混叠的图 3.17 作对比。条件(3.5.12)表示，对于最高频

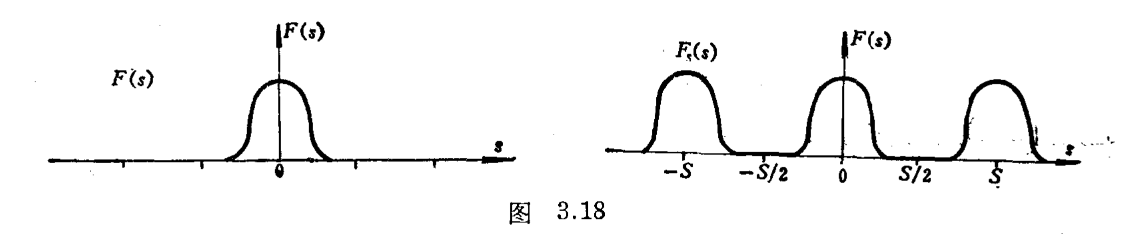


图 3.18

的一个正弦整波至少要有两个样点，才可以避免混叠。在此情况下，根据(3.5.13)，(3.5.10)，(3.5.4)有

$$f(t)=\int_{-\infty}^{\infty}F(s)e^{2\pi ist}ds=\int_{-\infty}^{\infty}\text{rect}s/S\cdot F_S(s)e^{2\pi ist}ds=\int_{-S/2}^{S/2}F_S(s)e^{2\pi ist}ds$$

$$=\int_{-S/2}^{S/2}\left[\Delta t\sum_{j=-\infty}^{\infty}f(j\Delta t)e^{-2\pi isj\Delta t}\right]e^{2\pi ist}ds=\sum_{j=-\infty}^{\infty}f(j\Delta t)\int_{-S/2}^{S/2}\Delta t\,e^{2\pi is(t-j\Delta t)}ds$$

$$=\sum_{j=-\infty}^{\infty}f(j\Delta t)\int_{-1/2\Delta t}^{1/2\Delta t}\Delta t\,e^{2\pi is(t-j\Delta t)}ds$$

算出右端的积分值，就得到

$$f(t)=\sum_{j=-\infty}^{\infty}f(j\Delta t)\text{sinc}(t-j\Delta t)/\Delta t,\quad -\infty<t<\infty \tag{3.5.14}$$

这就是说，对于满足(3.5.11)的有限谱宽函数$f(t)$，当取样间距满足$\Delta t\leqslant 1/2L$时，单凭样点值$f(j\Delta t)$，$j=0,\pm1,\pm2,\cdots$用公式(3.5.14)可以把函数$f(t)$完全复原。这一事实通常叫做抽样定理或取样定理，它有重要的实践意义。例如，在通信技术里，实际的信号波形总是具有有限谱宽的，或者近似于此，人们无需取其全部信息而只需按条件(3.5.12)取样就可能把信号完全复原。在涉及谐波分析的数值计算里，条件(3.5.12)对选取步长Δt也有指导意义。

当函数$f(t)$不具有有限谱宽时，混叠效应在原则上是不可避免的，但一般随Δt的缩小而减弱。

综合上述可知离散取样的两个参数Δt, T的作用是不同的，缩小Δt可以减少混叠但不能减少渗漏。放大T可以减少渗漏但不能减少混叠。

对于谱函数$F(s)$，根据(3.5.13)，(3.5.10)有

$$F(s)=\text{rect}\,s/S\cdot F_S(s)=\begin{cases}\Delta t\sum\limits_{j=-\infty}^{\infty}f(j\Delta t)e^{-2\pi ijs/S}, & |s|\leqslant S/2\\ 0, & |s|>S/2\end{cases} \tag{3.5.15}$$

这就是利用样点值$f(j\Delta t)$产生$F(s)$的公式。取整数N，使得$\Delta s=S/N$为要求辨认频率的"细度"，于是

$$F(k\Delta s)=\Delta t\sum_{j=-\infty}^{\infty}f(j\Delta t)e^{-2\pi ijk\Delta s/S}=\Delta t\sum_{j=-\infty}^{\infty}f(j\Delta t)e^{-2\pi ijk/N}$$

$$=\Delta t\sum_{j=0}^{N-1}\sum_{m=-\infty}^{\infty}f((j-mN)\Delta t)e^{-2\pi i(j-mN)k/N}$$

$$=\Delta t\sum_{j=0}^{N-1}\sum_{m=-\infty}^{\infty}f((j-mN)\Delta t)e^{-2\pi ijk/N}$$

命

$$F_k=F(k\Delta s) \tag{3.5.16}$$

$$f_j=\sum_{m=-\infty}^{\infty}f(j\Delta t-mT),\quad T=N\Delta t \tag{3.5.17}$$

于是 F_k, f_j 之间存在离散傅氏变换的关系

$$F_k = \Delta t \sum_{j=0}^{N-1} f_j W_N^{-jk}, \quad k=0, 1, \cdots, N-1 \tag{3.5.18}$$

注意用这套公式来计算谱点的值是准确的，用到了全部离散样点值 $f(j\Delta t)$，$j=0$，± 1，± 2，$\cdots$。

3.5.2 有限窗宽与频谱渗漏效应

实践上处理谐波分析时，人们不能掌握全部时间 $-\infty<t<\infty$ 而只能截取有限的时段，例如 $-T/2\leqslant t\leqslant T/2$ 来观察 $f(t)$，仿佛是通过一个有限宽的"窗口"来观察。这时

$$f(t), \quad -\infty<t<\infty \tag{3.5.19}$$

被代以

$$h(t)=g\left(\frac{t}{T}\right)\cdot f(t)=\begin{cases} f(t), & |t|\leqslant T/2 \\ 0, & \text{它处} \end{cases} \quad -\infty<t<\infty \tag{3.5.20}$$

如图 3.19。此处 $g(t)=\operatorname{rect} t$ 称为"数据窗"或截断函数，而

$$g\left(\frac{t}{T}\right)=\operatorname{rect} t/T \sim T\operatorname{sinc}(Ts)=\frac{\sin \pi Ts}{\pi s} \tag{3.5.21}$$

这就相当于 $f(t)$ 的谱函数 $F(s)$

被代以

$$H(s)=F(s)*T\operatorname{sinc}(Ts)=\int_{-\infty}^{\infty} F(\sigma)T\operatorname{sinc}(T(s-\sigma))d\sigma \tag{3.5.22}$$

它的效果是把 $F(s)$ 平滑化，从而破坏了频谱的精细结构。

典型的情况如取 $f(t)\equiv 1\sim F(s)=\delta(s)$，则 $H(s)=T\operatorname{sinc} Ts$，即一根无穷窄的谱线被拉宽或扩散成一个波形中间有高峰——这是近端干扰；两端有按 $\frac{1}{|s|}$ 衰减的正负"边瓣"，左右第一个零点在 $s=\pm\frac{1}{T}$——这是远端干扰。参看图 3.6(取 $T=1$)。注意 T 增大时 $g\left(\frac{t}{T}\right)$

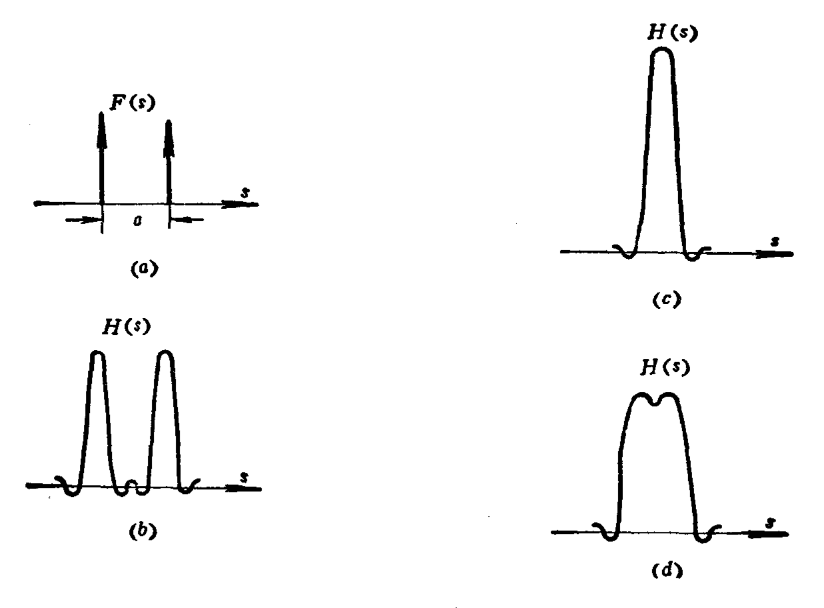


图 3.19

愈宽，$T\operatorname{sinc}Ts$愈窄；而 $T\to\infty$ 时，$g\left(\frac{t}{T}\right)\to 1$ 而 $T\operatorname{sinc}Ts\to\delta(s)$。

取 $f(t)=\cos\pi at\sim F(s)=\frac{1}{2}\delta\left(s+\frac{a}{2}\right)+\frac{1}{2}\delta\left(s-\frac{a}{2}\right)$，这是相距为 a 的两根谱线（图 3.19 的 a）。它被扩散为 $H(s)=\frac{T}{2}\operatorname{sinc}T\left(s+\frac{a}{2}\right)+\frac{T}{2}\operatorname{sinc}T\left(s-\frac{a}{2}\right)$。如取 $T\gg\frac{1}{a}$，则 $H(s)$ 作双峰状可以明确分辨（图 3.19 的 b）。如取 $T\ll\frac{1}{a}$，则 $H(s)$ 的两项汇合为单峰状，不能分辨（图 3.19 的 c）。从直观上看，临界分辨的情况是 $T\approx\frac{1}{a}$，这时一个 sinc 的主峰与另一个 sinc 的第一零点相重（图 3.19 的 d）。

这就是有限取样范围导致的频谱渗漏效应（leakage）。概言之，由于取样范围不能是无穷大而只能是有限大，因而频谱分辨率的精细度不能是无穷小而只能是有限小。

显然当 $f(t)$ 本身具有有限宽度，即当 $|t|\geqslant A$，$f(t)\equiv 0$ 时，只需取 $T\gg 2A$ 就没有渗漏效应；在其它情况下渗漏是不能避免的，但其效应随窗宽 T 的增大而减弱。

为了减低渗漏干扰，特别是远端干扰，还可以取另外的"数据窗"。为了比较，连同(3.5.20)列举一些实用上可取的截断函数及其谱函数如表 3.4。图形见图 3.20、图 3.21。这里 g_0 是矩形；g_1 是三角形；g_2 是截取余弦函数的一段，有导数的连续性；g_3 形状接近于矩形，但把棱角修匀，有二阶导数的连续性。

表 3.4

$g(t)$	$G(s)$
$g_0(t)=\operatorname{rect} t$	$G_0(s)=\frac{\sin\pi s}{\pi s}\approx O(\lvert s\rvert^{-1})$
$g_1(t)=\Lambda(2t)$	$G_1(s)=\frac{1}{2}\left(\frac{\sin\pi s/2}{\pi s/2}\right)^2\approx O(\lvert s\rvert^{-2})$
$g_2(t)=\operatorname{rect} t\cdot\frac{1}{2}(1+\cos 2\pi t)$	$G_2(s)=\frac{1}{2}\,\frac{\sin\pi s}{\pi s(1-s^2)}\approx O(\lvert s\rvert^{-3})$
$g_3(t)=\begin{cases}1, \lvert t\rvert\leqslant 0.45\\ \frac{1}{2}(1+\cos 2\pi(10t-4.5)), \quad 0.45\leqslant\lvert t\rvert\leqslant 0.5\\ 0, \lvert t\rvert>0.5\end{cases}$	$G_3(s)=\frac{\sin\pi s+\sin\frac{9\pi s}{10}}{2\pi s(1-s^2/100)}\approx O(\lvert s\rvert^{-3})$

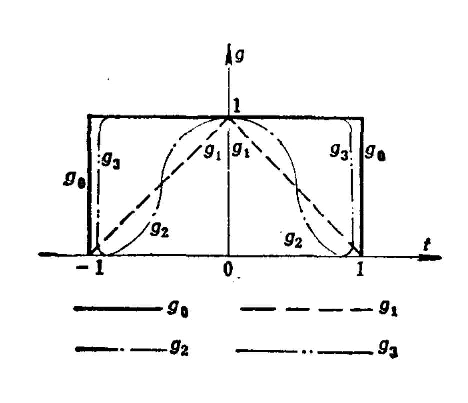


图 3.20

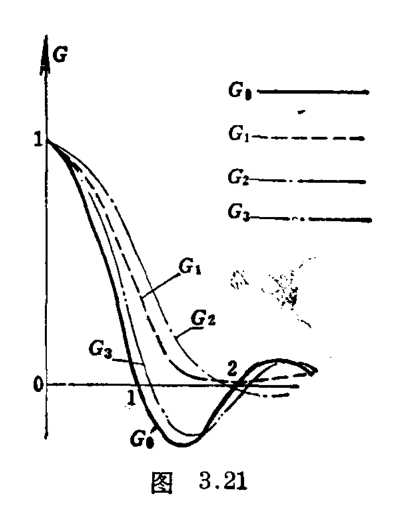


图 3.21

所有这些截断函数当然都满足

1. $g(t)\equiv 0$, 当 $|t|\geqslant\frac{1}{2}$;

2. $g\left(\frac{t}{T}\right)\to 1$, 即 $TG(Ts)\to\delta(s)$, 当 $T\to\infty$。

从表 3.4 及图 3.20、3.21 可以看出，若 $g(t)$ 愈“方”，则 $G(s)$ 的主峰愈窄，近端干扰较小，最好的就是 G_0。反之 $g(t)$ 愈光滑，则 $G(s)$ 的“边瓣”衰退愈快，远端干扰较小，如 $g_2\sim G_2$。G_3 的图形在近端接近于 G_0，远端接近于 G_2，因此是一个折衷的选取。注意，取 $g=g_0$ 时，事实上对数据无需作预先加工，而取其它的 g 时要作预先的乘法处理。

3.5.3 连续与离散傅氏变换的关系

设有一对傅氏变换

$$f(t)\sim F(s)$$

取五个正参数 N, Δt, Δs, T, S 满足下列四个关系

$$T=N\Delta t,\quad S=N\Delta s,\quad \Delta t=\frac{1}{S},\quad \Delta s=\frac{1}{T} \tag{3.5.23}$$

对 f, F 分别构造周期复变函数

$$\begin{cases} f_T(t)=\sum\limits_{m=-\infty}^{\infty} f(t-mT) \\ F_S(s)=\sum\limits_{m=-\infty}^{\infty} F(s-mS) \end{cases} \tag{3.5.24}$$

这里，自然假定 $f(t)$ 及 $F(s)$ 在无穷远处衰减得足够快，以使无穷和收敛。显然 $f_T(t)$, $F_S(s)$ 各具周期 T, S

$$f_T(s+T)\equiv f(s),\quad F_S(s+S)=F_S(s) \tag{3.5.25}$$

命

$$\begin{cases} u_j=\Delta t f_T(j\Delta t)=\Delta t\sum\limits_{m=-\infty}^{\infty} f(j\Delta t-mT),\quad j=0,\ \pm1,\ \pm2,\ \cdots \\ U_k=F_S(k\Delta s)=\sum\limits_{m=-\infty}^{\infty} F_S(k\Delta s-mS),\quad k=0,\ \pm1,\ \pm2,\ \cdots \end{cases} \tag{3.5.26}$$

显然 $\{u_j\}$ 与 $\{U_k\}$ 都是周期为 N 的序列

$$u_j\equiv u_{j+N},\quad U_k\equiv U_{k+N} \tag{3.5.27}$$

因此各取 N 个分量如 $(u_0,\ u_1,\ \cdots,u_{N-1})$, $(U_0,U_2,\cdots,U_{N-1})$ 就能分别决定全序列 $\{u_j\}$, $\{U_k\}$。重要的事实在于这样的 (u_j) 和 (U_k) 构成一对离散傅氏变换，即

$$U_k=\sum_{j=0}^{N-1} u_j W_N^{-jk}\qquad k=0,\ \cdots,\ N-1 \tag{3.5.28}$$

$$u_j=\frac{1}{N}\sum_{k=0}^{N-1} U_k W_N^{jk}\qquad j=0,\ \cdots,\ N-1 \tag{3.5.29}$$

事实上，由于函数 $f_T(t)$ 具有周期 T，故可展为傅氏级数

$$f_T(t)=\sum_{k=-\infty}^{\infty} c_k e^{2\pi ikt/T}\quad -\infty<t<\infty$$

$$
\begin{aligned}
c_k &= \frac{1}{T}\int_0^T f_T(t)e^{-2\pi ikt/T}dt = \frac{1}{T}\int_0^T \sum_{m=-\infty}^{\infty} f(t-mT)e^{-2\pi ikt/T}dt \\
&= \frac{1}{T}\int_0^T \sum_{m=-\infty}^{\infty} f(t-mT)e^{2\pi ik(t-mT)/T}dt = \frac{1}{T}\sum_{m=-\infty}^{\infty}\int_{mT}^{(m+1)T} f(t)e^{-2\pi ikt/T}dt \\
&= \frac{1}{T}\int_{-\infty}^{\infty} f(t)e^{-2\pi ikt/T}dt = \frac{1}{T}\int_{-\infty}^{\infty} f(t)e^{-2\pi itk\Delta s}dt = \frac{1}{T}F(k\Delta s)
\end{aligned}
$$

因此

$$
f_T(t) = \frac{1}{T}\sum_{k=-\infty}^{\infty} F(k\Delta s)e^{2\pi ikt/T}
$$

$$
\begin{aligned}
u_j &= \Delta t f_T(j\Delta t) = \frac{\Delta t}{T}\sum_{k'=-\infty}^{\infty} F(k'\Delta s)e^{2\pi ijk'\Delta t/T} = \frac{1}{N}\sum_{k'=\infty}^{\infty} F(k'\Delta s)e^{2\pi ijk'/N} \\
&= \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=-\infty}^{\infty} F((k-mN)\Delta s)\, e^{2\pi ij(k-mN)/N} = \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=-\infty}^{\infty} F(k\Delta s-mS)\,e^{2\pi ijk/N} \\
&= \frac{1}{N}\sum_{k=0}^{N-1} F_S(k\Delta s)e^{2\pi ijk/N} = \frac{1}{N}\sum_{k=0}^{N-1} U_k W_N^{jk}
\end{aligned}
$$

故关系(3.5.28)，(3.5.29)成立。关系式(3.5.28)，(3.5.29)可以更明显地表成

$$
F_S(k\Delta s) = \Delta t \sum_{j=-N'}^{-N'+N-1} f_T(j\Delta t)W_N^{-jk} \qquad k=-N', \cdots, -N'+N-1 \tag{3.5.30}
$$

$$
f_T(j\Delta t) = \Delta s \sum_{k=-N'}^{-N'+N-1} F_S(k\Delta s)W_N^{jk} \qquad j=-N', \cdots, -N'+N-1 \tag{3.5.31}
$$

考虑 $f(t)$ 为有限谱宽函数的特例。这时存在 $S>0$ 使得

$$
F(s)\equiv 0 \quad 当 \quad |s|\geqslant S/2
$$

因此有(图 3.18)

$$
F_S(s)\equiv F(s) \quad 当 \quad |s|\leqslant S/2 \tag{3.5.32}
$$

对于任意整数 $N>0$，取 $\Delta s=S/N$，$\Delta t=1/S$，$T=N\Delta t$，则关系式 (3.5.28) 得到满足，于是由(3.5.26)以及(3.5.30)可得谱函数 $F(s)$ 在离散点 $s=k\Delta s$ 的准确表达式

$$
\begin{cases}
F(k\Delta s) = \sum\limits_{j=0}^{N-1} u_j W_N^{-jk}, \quad |k\Delta s|\leqslant S/2 \\
u_j = \Delta t f_T(j\Delta t) = \Delta t \sum\limits_{m=-\infty}^{\infty} f(j\Delta t - mT)
\end{cases} \tag{3.5.33}
$$

§ 3.6 谱的近似计算

下面简单介绍快速傅氏变换的个别应用。有关的问题和其它的应用可以参考[3]。

3.6.1 傅氏级数的近似计算

设函数 $f(t)$ 具有周期 T，它可展为傅氏级数

$$
f(t) = \sum_{k=-\infty}^{\infty} c_k e^{2\pi ikt/T}, \quad -\infty<t<\infty \tag{3.6.1}
$$

现将基本周期 $[0, T]$ 分为 N 等分，要求在离散点值

$$
f_j = f(j\Delta t), \quad j=0, \cdots, N-1, \quad \Delta t = T/N \tag{3.6.2}
$$

的基础上计算傅氏系数 c_k。由于 f 是周期函数，自然采用矩形求积公式⊖ (见第二章)

⊖ 当端点值 $f(0)$，$f(T)$ 不相等时，$f(t)$ 作为周期函数是间断的，这时取 $f_0=\frac{1}{2}(f(0)+f(T))$，而矩形公式实质上就是梯形公式。

$$c_k=\frac{1}{T}\int_0^T f(t)e^{-2\pi ikt/T}dt\sim\frac{\Delta t}{T}\sum_{j=0}^{N-1}f(j\Delta t)e^{-2\pi ikj\Delta t/T}=\frac{1}{N}\sum_{j=0}^{N-1}f_jW_N^{-jk}$$

$$k=0,\ \pm1,\ \pm2,\ \cdots \tag{3.6.3}$$

而 $f(t)$ 则近似地表为

$$f(t)\sim\sum_k c'_k e^{2\pi ikt/T},\qquad -\infty<t<\infty \tag{3.6.4}$$

关于右端取多少项，取哪些项的问题留待稍后再说。由于

$$c'_{k+N}=\frac{1}{N}\sum_{j=0}^{N-1}f_jW_N^{-j(k+N)}=\frac{1}{N}\sum_{j=0}^{N-1}f_jW_N^{-jk}=c'_k,\quad k=0,\ \pm1,\ \pm2,\ \cdots$$

故 $\{c'_k\}$ 为周期序列，周期为 N，可以由任意相连的 N 个分量例如 $\{c'_0,\ c'_1,\ \cdots,\ c'_{N-1}\}$ 完全决定。由式(3.6.3)可知 $\{Nc'_0,\ \cdots,\ Nc'_{N-1}\}$ 与 $\{f_0,\ \cdots,\ f_{N-1}\}$ 为一对互逆的离散傅氏变换

$$Nc'_k=\sum_{j=0}^{N-1}f_jW_N^{-jk},\quad k=0,\ \cdots,\ N-1 \tag{3.6.5}$$

$$f_j=\sum_{k=0}^{N-1}c'_kW_N^{jk},\quad j=0,\ \cdots,\ N-1 \tag{3.6.6}$$

可以用快速算法来计算 c'_k。

近似系数 $\{c'_k\}$ 与真系数 $\{c_k\}$ 之间有简单的关系。为此，将 $t=j\Delta t$ 代入(3.6.1)得到

$$f_j=f(j\Delta t)=\sum_{k'=-\infty}^{\infty}c_{k'}e^{2\pi ijk'\Delta t/T}=\sum_{k'=-\infty}^{\infty}c_{k'}W_N^{jk'}$$

$$=\sum_{k=0}^{N-1}\sum_{m=-\infty}^{\infty}c_{k-mN}W_N^{j(k-mN)}=\sum_{k=0}^{N-1}\left(\sum_{m=-\infty}^{\infty}c_{k-mN}\right)W_N^{jk}$$

与(3.6.6)比较即得

$$c'_k=\sum_{m=-\infty}^{\infty}c_{k-mN},\quad k=0,\ \pm1,\ \pm2,\ \cdots \tag{3.6.7}$$

从而得到误差

$$c'_k-c_k=\sum_{\substack{m=-\infty\\ m\neq0}}^{\infty}c_{k-mN} \tag{3.6.8}$$

这就是说，近似列 $\{c'_k\}$ 可由原列 $\{c_k\}$ 左右平移 0，$\pm N$，$\pm2N,\cdots$ 再叠加而得。这里显示了离散取样的频率混叠效应。在离散点值(3.6.2)的基础上不能辨认 $|k|\geqslant N/2$ 以上的频率，原有高频分量都被折叠到 $|k|\leqslant N/2$ 的范围。一般地，当 $f(t)$ 的光滑性提高时，c_k 随 $k\to\pm\infty$ 的衰减率愈快。图 3.22 中反映了一个函数的傅氏系数 c_k(圆点)和近似系数 c'_k(叉点)。注意，图 3.22 与图 3.17 是相似的。

现在转来讨论表达式(3.6.4)中取多少项以及取哪些项的问题。首先，表达式中的项数

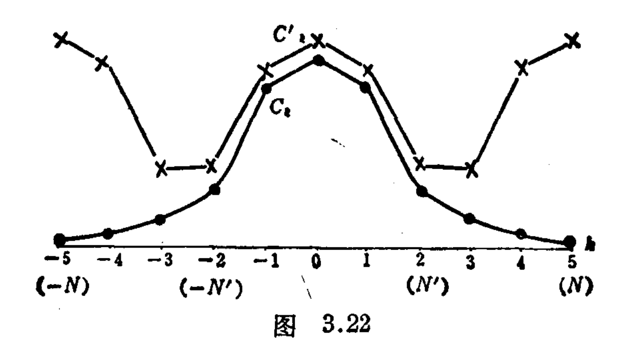


图 3.22

决不是愈多愈好，比如说，取 $k=-\infty$ 至 $+\infty$，则由于 $\{c'_k\}$ 的周期性，级数是发散的。由于 $\{c'_k\}$ 中只有 N 个相连独立分量是独立的，因此自然地应取连续的 N 个项 $k=m, m+1, \cdots, m+N-1$。从公式(3.6.8)或图 3.22 可以看出，取 $[-N/2, N/2]$ 内 N 个项与原系数的偏差为最小，可以从图 3.22 比较 $[-N/2, N/2]$ 与 $[0, N]$ 的情况。因此作为合理的逼近应取

$$f(t)\approx\sum_{|k|\leqslant N/2}c'_k e^{2\pi ikt/T}=\sum_{k=-N'}^{N'}c'_k e^{2\pi ikt/T} \tag{3.6.9}$$

$N=2N'$ 或 $2N'+1$，可以证明，当 $f(t)$ 有连续的 $r(r\geqslant 2)$ 阶导数，各阶导数均有周期 T，$N=2N'+1$ 则有误差估计式[4]：

$$\left|f(t)-\sum_{k=-N'}^{N'}c'_k e^{2\pi ikt/T}\right|\leqslant K_r M_r/N'^{r-1} \tag{3.6.10}$$

此处

$$M_r=\max|f^{(r)}(t)|,\quad K_r=\frac{4}{r-1}\left(\frac{T}{2\pi}\right)^r$$

由于周期函数在一个基本周期 $[0, T]$ 上的行为足以反映它在 $(-\infty, \infty)$ 上的全貌，因此取样(3.6.2)虽然局限于有穷的范围 $[0, T]$ 内，但并不引起频谱渗漏效应。

3.6.2 谱函数的近似计算

在 $-\infty<t<\infty$ 上的函数 $f(t)$ 可以看作是周期函数当周期 $T\to\infty$ 时的极限，而谱函数 $F(s)$ 则可看作是相应的傅氏系数的极限。据此可以制定从 $f(t)$ 的离散点值计算 $F(s)$ 的离散点值的计算方法如下：

1. 适当估计待定谱函数 $F(s)$，特别是根据 $f(t)$ 的光滑度来估计 $F(s)$ 在 $s=\pm\infty$ 处的衰减率。据此选取足够小的 Δt，使得区间 $|s|\leqslant S/2$ 足以包括谱的主要部分。此处 $S=1/\Delta t$。

也可以反过来说，当原函数本身是以 Δt 为间距的离散数据形式给出时，则能够计算谱的范围只是 $|s|\leqslant S/2$。

2. 取足够大的整数 N 使得 $S/N=\Delta s$ 为所要求的频谱精细分辨的下限，于是相应地原函数 f 的"数据窗"的宽度应取为 $T=N\Delta t=1/\Delta s$，即取样范围为 $-T/2\leqslant t\leqslant T/2$。

也可以反过来说，当所掌握的原函数数据范围为 $-T/2\leqslant t\leqslant T/2$ 时，则频谱精细分辨的下限只是 $\Delta s=1/T=S/N$。

这样，问题就是在范围 $|t|\leqslant T/2$ 内 N 个等距离散数据 $f(j\Delta t)$ 的基础上计算在范围 $|s|\leqslant S/2$ 内 N 个等距频率点值 $F(k\Delta s)$。这里有五个参数 $\Delta t, T, \Delta s, S, N$，满足关系式

$$\Delta t=T/N,\quad \Delta s=S/N,\quad S=1/\Delta t,\quad T=1/\Delta s \tag{3.6.11}$$

其中有三个独立的关系式，因此五个参数中只有两个是独立的。

3. 把傅氏积分的无穷限 $\pm\infty$ 代以有穷限 $\pm T/2$，并运用矩形式求积公式

$$F(k\Delta s)=\int_{-\infty}^{\infty}f(t)e^{-2\pi itk\Delta s}dt\approx\int_{-T/2}^{T/2}f(t)e^{-2\pi itk\Delta s}dt\approx\Delta t\sum_{j=-m}^{-m+N-1}f(j\Delta t)e^{-2\pi ij\Delta tk\Delta s}$$

$$k=-N', -N'+1, \cdots, -N'+N-1$$

$$N=2N' \quad 或 \quad 2N'+1$$

即

$$F(k\Delta s)\approx F_k=\Delta t\sum_{j=-m}^{-m+N-1}f_j W_N^{-jk}\qquad k=-N', \cdots, -N'+N-1 \tag{3.6.12}$$

$$f_j=f(j\Delta t),\ j=-N',\ \cdots,\ -N'+N-1$$㊀

即 $\{\Delta t f_j\}$ 与 $\{F_k\}$ 之间为一对离散傅氏变换。各自对下标 j, k 按周期 N 进行延拓就归化为标准型，即 $f_{-j}=f_{N-j}$, $F_{-k}=F_{N-k}$。

$$F_k=\Delta t\sum_{j=0}^{N-1}f_jW_N^{-jk}\qquad k=0,\ \cdots,\ N-1 \tag{3.6.13}$$

而谱函数的估值即由(3.6.12)给出。

由于 Δt 是有限小，T 是有限大，一般地总有频谱的混叠和渗漏。如果面临的任务是计算"数学函数"的谱，即函数 $f(t)$ 在 $(-\infty,\ \infty)$ 上有明显表达式，则可以用下列办法来克服渗漏。代替(3.6.2)改取

$$f_j=f_T(j\Delta t)=\sum_{m=-\infty}^{\infty}f(j\Delta t-mT),\quad j=0,\ \cdots,\ N-1 \tag{3.6.14}$$

通过离散变换(3.6.13)得到 F_k, $k=0,\ \cdots,\ N-1$ 再用关系

$$F_{-k}=F_{N-k} \tag{3.6.15}$$

归化区间 $|k|\leqslant N/2$。于是根据(3.5.24)可知准确地表为，

$$F_k=F_S(k\Delta s)=\sum_{m=-\infty}^{\infty}F(k\Delta s-mS),\quad |k|\leqslant N/2\quad 即\quad |k\Delta s|\leqslant S/2 \tag{3.6.16}$$

因此所得结果不是谱函数本身而是它的周期性重复。因此误差为(见图 3.17)

$$F_k-F(k\Delta s)=\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}F(k\Delta s-mS),\quad |k\Delta s|\leqslant S/2 \tag{3.6.17}$$

这单纯是由混叠效应引起的，随着 Δt 的减小即 S 的增大而减小。如果 $f(t)$ 是有限频谱函数，则取 Δt 充分小即 S 充分大时可以完全消除误差，如在 3.5.3 节末所述。

如果原函数 $f(t)$ 并不是明确定义的数学函数而只是某个观测序列的离散值，它们当然只能是有限长。设"数据窗"为 $-T/2\leqslant t\leqslant T/2$。这就是如 3.5.2 节中所述的，取

$$f_j=f(j\Delta t)$$

这相当于把原函数 $f(t)$ 代以 $h(t)=f(t)\cdot g(t)$，即谱函数 $F(s)$ 代以 $H(s)=F(s)*G(s)$，而

$$g(t)=\operatorname{rect}\frac{t}{T}=\begin{cases}1, & |t|<T/2\\ \dfrac{1}{2}, & |t|=T/2\\ 0, & |t|>T/2\end{cases}$$

$$G(s)=T\operatorname{sinc}(Ts)=\frac{\sin(\pi Ts)}{\pi s}$$

于是由于

$$h_T(t)\equiv h(t)\equiv f(t),\quad |t|\leqslant T/2$$

以及公式(3.5.30)，

$$H_S(k\Delta s)=\sum_{j=-N'}^{-N'+N-1}h_T(j\Delta t)W_N^{-jk}=\sum_{j=-N'}^{-N'+N-1}h(j\Delta t)W_N^{-jk}=\sum_{j=-N'}^{-N'+N-1}f(j\Delta t)W_N^{-jk}$$

因此所得的离散傅氏变换 F_k(归化到 $|k\Delta s|\leqslant S/2$ 上)

$$F_k=H_S(k\Delta s),\quad |k\Delta s|\leqslant S/2$$

即不是谱函数 $F(s)$ 的值，而是 $f(t)\cdot g(t)$ 的谱函数 $H(s)$ 的周期性重复。因此，这里除了有

㊀ 当 $f(T/2)\neq f(-T/2)$ 时，应取 $f_0=\frac{1}{2}(f(T/2)+f(-T/2))$，事实上相当于采用梯形公式。

频率混叠外，还有有限窗宽带来的渗漏效应。关于后者的讨论和改进办法已见 3.5.2 节。

3.6.3 功率谱的估算

许多工程技术问题中需要估计时间序列的功率谱。对于连续的实数时间序列 $f(t)$，$-\infty<t<\infty$，功率谱通常定义为

$$P(s)=\lim_{T\to\infty}\frac{2}{T}\left|\int_{-T/2}^{T/2}f(t)e^{-2\pi ist}dt\right|^2 \tag{3.6.18}$$

取

$$g(t)=\operatorname{rect}t=\begin{cases}1, & \text{当}\quad |t|\leqslant\dfrac{1}{2}\\[2mm] 0, & \text{当}\quad |t|>\dfrac{1}{2}\end{cases} \tag{3.6.19}$$

即

$$g(t/T)=\operatorname{rect}(t/T)=\begin{cases}1, & \text{当}\quad |t|\leqslant T/2\\ 0, & \text{当}\quad |t|>T/2\end{cases}$$

于是

$$\int_{-T/2}^{T/2}f(t)e^{-2\pi ist}dt=\int_{-\infty}^{\infty}f(t)g(t/T)e^{-2\pi ist}dt$$

$g(t/T)$ 就是对应于窗口 $[-T/2,\ T/2]$ 的“数据窗”或截断函数，见 3.5.2 节。因此

$$P(s)=\lim_{T\to\infty}\frac{2}{T}\left|\int_{-\infty}^{\infty}f(t)g(t/T)e^{-2\pi ist}dt\right|^2 \tag{3.6.20}$$

通常只在有穷的区间 $[-T/2,\ T/2]$ 内掌握 $f(t)$ 的数据，故作为 $P(s)$ 的估计，可以取

$$P(s)\approx\frac{2}{T}\left|\int_{-\infty}^{\infty}f(t)g(t/T)e^{-2\pi ist}dt\right|^2 \tag{3.6.21}$$

问题归结于计算 $f(t)g(t/T)$ 的谱函数。

由于取样在有限的范围，必有渗漏。为了减少渗漏，特别是谱线扩散的远端干扰，可以选取比式(3.6.19)更好一些的“数据窗”，例如取 $g(t)=g_3(t)$ (3.5.2 节)。

不妨将函数 $f(t)g(t/T)$ 作平移 $t\to t+T/2$，即变为 $f(t+T/2)g\left(\dfrac{t+T/2}{T}\right)$，原函数经平移后相当于对谱函数乘以绝对值为 1 的因子 $e^{2\pi isT/2}$，由于功率谱定义为谱函数的绝对值平方，故 $P(s)$ 不变。对 $f\cdot g$ 作平移相当于以数据 f 的起始端($t=0$)作为时间的原点。即把窗口 $[-T/2,\ T/2]$ 移为 $[0,\ T]$，故不妨就认为：$f(t)$ 定义在 $[0,\ T]$ 上，而 $g(t/T)$ 为对应于窗口 $[0,\ T]$ 的“数据窗”，于是

$$P(s)\approx\frac{2}{T}\left|\int_{-\infty}^{\infty}f(t)g(t/T)e^{-2\pi ist}dt\right|^2=\frac{2}{T}\left|\int_{0}^{T}f(t)g(t/T)e^{-2\pi ist}dt\right|^2 \tag{3.6.22}$$

如果取 $g=g_3$(见 3.5.2 节)则对于窗口 $[0,\ T]$ 的表达式是

$$g(t/T)=g_3(t/T)=\begin{cases}0 & t\leqslant 0\\ \dfrac{1}{2}\left\{1-\cos\left(20\pi\dfrac{t}{T}\right)\right\} & 0\leqslant t\leqslant 0.05T\\ 1 & 0.05T\leqslant t\leqslant 0.95T\\ \dfrac{1}{2}\left\{1-\cos\left(20\pi\dfrac{t}{T}\right)\right\} & 0.95T\leqslant t\leqslant T\\ 0 & T\leqslant t\end{cases} \tag{3.6.23}$$

因此，如命 $\{U_k\}$ 为 $\{u_j\}$ 的离散傅氏变换

$$U_k=\sum_{j=0}^{N-1}u_jW_N^{-jk},\quad k=0,\ 1,\ \cdots,\ N-1$$

$$U_{-k}=U_{N-k},\quad k=0,\ 1,\ \cdots,\ N-1$$

则有

$$U(k\Delta s)=\Delta tU_k,\quad -N/2\leqslant k\leqslant N/2$$

而功率谱则为

$$P(k\Delta s)\approx\frac{2(\Delta t)^2}{T}|U_k|^2=\frac{2\Delta t}{N}|U_k|^2 \tag{3.6.24}$$

k 只取一半：$k=0,\ 1,\ \cdots,\ N/2$，即 $0\leqslant k\Delta s\leqslant S/2$。

对于快速傅氏变换，为了提高方法的效率和编程序方便，通常是取 $N=2^m$，即 N 为 2 的整次幂。为此，对于原来的时间序列 $\{f_j\}$，在必要时需要截去一段或者补加一些 0，以使长度 N 成为 2 的整次幂。

功率谱的估算过程大致归纳如下：

1．对原始数据截去一段尾巴或补加一些 0，使得序列长度 $N=2^m$，由于

$$u(t)\equiv f(t)\cdot g(t/T) \tag{3.6.25}$$

为实函数，它的谱函数

$$U(s)=\int_{-\infty}^{\infty}u(t)e^{-2\pi ist}dt \tag{3.6.26}$$

为共轭对称，$U(-s)=U^*(s)$，而 $P(s)$ 是按照绝对值定义的故为对称，即 $P(-s)=P(s)$，因此只须考虑 $s\geqslant0$，而

$$P(s)\approx\frac{2}{T}|U(s)|^2 \tag{3.6.27}$$

实践上，时间序列 $f(t)$ 通常给为等距离散样本的形式

$$f_j=f(j\Delta t),\quad j=0,\ 1,\ \cdots,\ N-1 \tag{3.6.28}$$

命 $T=N\Delta t$，$S=1/\Delta t$，$\Delta s=S/N$。可以按照 3.6.2 节的方法在

$$u_j=f(j\Delta t)g(j\Delta t/T),\quad j=0,\ 1,\ \cdots,\ N-1 \tag{3.6.29}$$

的基础上计算

$$U(k\Delta s)=\int_0^T u(t)e^{-2\pi ik\Delta st}dt\approx\Delta t\sum_{j=0}^{N-1}u_je^{-2\pi ij\Delta tk\Delta s}$$

$$-N/2\leqslant k\leqslant N/2\quad 即\quad |k\Delta s|\leqslant S/2$$

$$f_j=f(j\Delta t),\quad j=0,\ 1,\ \cdots,\ N-1,\quad T=N\Delta t$$

2．使用适当的"数据窗"$g(t/T)$ 对 f_j 进行修正

$$u_j=f_jg_j,\quad j=0,\ 1,\ \cdots,\ N-1$$

此处 g 可取为 g_3(3.6.23)或其它；

$$g_j=g(j\Delta t/T)$$

$$u_j=f_jg_j,\quad j=0,\ 1,\ \cdots,\ N-1$$

3．快速算法计算实数列 $\{u_j\}$ 的离散傅氏变换

$$U_k=\sum_{j=0}^{N-1}u_jW_N^{-jk},\quad k=0,\ 1,\ \cdots,\ N-1$$

4．得到功率谱估计

$$P(k\Delta s)\approx\frac{2\Delta t}{N}|U_k|^2,\quad k=0,\ 1,\ \cdots,\ N/2$$

此处 $\Delta s=1/N\Delta t=1/T$。

参 考 资 料

[1] 莱特希尔,《傅里叶变换与广义函数》,科学出版社, 1965。

[2] Cooley, Tukey, "An algorithm for the machine calculation of complex Fourier series", Math. of Computation, Vol. 19, No. 90(1965), 297～301.

[3] IEEE Trans, "Audio and Electroacoustics", Vol. AU-15, No. 3(1967);Vol. AU-17, No. 3(1969).

[4] 华罗庚,王元,《数值积分及其应用》§8,科学出版社, 1963。

第四章　曲线拟合与经验公式

§4.1　问题的提出

先看一个最简单的例子。

设 $x-y$ 平面上有五个点，其分布图形和观测数据 (x_k, y_k)，$k=1, 2, \cdots, 5$，如图 4.1 和表 4.1 所示。问题是要用一简单式子表示这些点的关系。由图易知，这些点大体分布在一直线上，因此可用线性式表为

$$y=a+bx \tag{4.1.1}$$

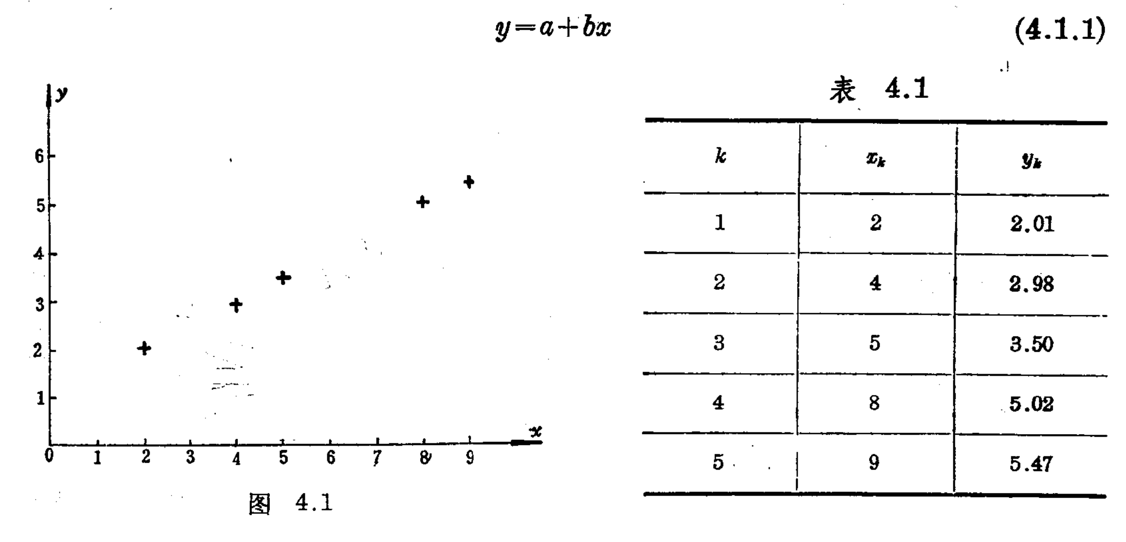


图　4.1

表　4.1

k	x_k	y_k
1	2	2.01
2	4	2.98
3	5	3.50
4	8	5.02
5	9	5.47

这时各观测数据大体满足如下方程组：

$$\begin{cases} a+2b=2.01 \\ a+4b=2.98 \\ a+5b=3.50 \\ a+8b=5.02 \\ a+9b=5.47 \end{cases} \tag{4.1.2}$$

式中的 a、b 是待定参数。这时问题简化为确定参数 a、b。

确定 a、b 的最简单方法是“选点法”。这时在给定的五个点中，任选两个连一直线。换句话说，从上述五个式子中任选两个联立便可解出 a、b。在实际问题中，由于给出的观测数据往往带有误差，甚至表达式本身也是近似的，因此选点法确定的直线一般不会同时通过给出的所有点，亦即这样确定的 a、b 不会使上述五个式子同时成立。而且随着点的选取不同，a、b 的值也有差异。例如选第 1、2 两点得

$$\begin{cases} a+2b=2.01 \\ a+4b=2.98 \end{cases} \quad \text{有解} \begin{cases} a=1.0400 \\ b=0.4850 \end{cases}$$

选第 2、3 两点得

$$\begin{cases} a+4b=2.98 \\ a+5b=3.50 \end{cases} \quad 有解 \begin{cases} a=0.9000 \\ b=0.5200 \end{cases}$$

为减小解的变化，可用“平均法”，即把上述五个式子分为两组，并分别求平均，最后从这两个平均式中解出 a、b。例如把前三式和后两式分别作为一组得

$$\begin{cases} a+3.7b=2.83 \\ a+8.5b=5.245 \end{cases} \quad 有解 \begin{cases} a=0.9748 \\ b=0.5014 \end{cases}$$

显然，平均法的解也随着分组的不同而有差异。例如改令第 1、2、4 式为一组，其它式为另一组则有

$$\begin{cases} a+4.6667b=3.3367 \\ a+7b=4.4850 \end{cases} \quad 有解 \begin{cases} a=1.0403 \\ b=0.4921 \end{cases}$$

总之，在实际问题中，只要给出的观测点数大于待定参数的个数，这时列出的方程组就会出现互相矛盾的现象。曲线拟合中参数的确定问题，实质上就是解矛盾方程组的问题。选点法与平均法只是解矛盾方程组中最初等的方法。因为它们的解都不唯一，而且在不同的解中，什么样的解“最好”以及怎样才能找到最好的解等问题尚不明确，因此在一般情况下，这两个方法仅作粗略估值之用，而不作为正式算法。

为使问题明确和便于研究，先建立待定参数好坏的标准。

在上述的例子中，如果有某一种方法可以确定参数 a、b，则给出 x 后便可算 y，记作

$$y_k^*=a+bx_k \qquad (k=1,\ 2,\ \cdots,\ 5) \tag{4.1.3}$$

这里的 y_k^* 称为 y_k 的估计值。如上所述，由于数据的误差和表达式的不精确等原因，估计值 y_k^* 与观测值 y_k 是不完全相同的，它们之间的差常称为“残差”或“剩余”，记为

$$e_k=y_k-y_k^*=y_k-(a+bx_k) \qquad (k=1,\ 2,\ \cdots,\ 5) \tag{4.1.4}$$

由于在原始数据给定的情况下，残差仅依赖于参数 a、b 的取值，因此残差的大小就是衡量被确定的参数 a、b 好坏的重要标志。

可以使用各种各样的原则来确定参数，例如：

(1) 使残差的最大绝对值达到最小，即令

$$T=\max_k |e_k| \qquad 最小$$

(2) 使残差的绝对值之和达到最小，即令

$$A=\sum_k |e_k| \qquad 最小$$

(3) 使残差的平方和达到最小，即令

$$Q=\sum_k e_k^2 \qquad 最小$$

显然，单独使用残差和 $\sum_k e_k$ 为最小的原则并不充分，因为即使 $\sum_k e_k=0$，也不能避免某些残差的绝对值很大的情况。(1)、(2) 两个原则是较自然的，也是较理想的，但计算繁难，远非原则 (3) 来得简单和常用。基于原则 (3) 得到的参数 a、b，通常称为最小二乘解。本章的重点就是在最小二乘意义下讨论曲线拟合和经验公式中参数的确定，至于非最小二乘问题，在本章的最后一节将稍有涉及。

曲线拟合问题的特点在于，被确定的直线（或曲线）原则上并不特别要求真正通过给定的某一个（或一些）点，而只要求它尽可能从给定点的附近通过。插值法所确定的曲线要求

通过给出的所有点。对于含有观测误差的数据来说，不过点的原则显然更为适合。因为这样的处理，可以部分地抵消数据中含有的观测误差。

最后，我们指出，确定表达式中的参数并不是曲线拟合问题的全部。曲线拟合中，首先碰到的问题是表达式形式的确定。这是参数估值的基础。但是，这一部分与客观实际联系非常紧密，必须深入实际作调查研究才能得到较好的解决。然而某些数学方法有可能被用来协助解决这类问题，在§4.4中将对此作较详细的讨论。此外，用"样条函数"(spline function)作曲线的分段拟合是国际上近十年来值得注意的动向，由于第一章已有详细介绍，这里就不介绍了。

§4.2 线性模型中参数的确定

4.2.1 基本算法

曲线拟合和经验公式中参数的最小二乘估计问题，可用如下的数学语言描述：

"若 y 是关于自变量 $\boldsymbol{X}$ 和待定参数 $\boldsymbol{B}$ 的形式已知的函数：

$$y=f(\boldsymbol{X}, \boldsymbol{B}) \tag{4.2.1}$$

今给出 $(\boldsymbol{X}, y)$ 的 n 对观测值：

$$(\boldsymbol{X}_k, y_k) \qquad (k=1, 2, \cdots, n) \tag{4.2.2}$$

要求确定参数 $\boldsymbol{B}$ 使

$$Q=\sum_{k=1}^{n}[y_k-f(\boldsymbol{X}_k, \boldsymbol{B})]^2 \tag{4.2.3}$$

为最小"。

这里的 $\boldsymbol{X}$ 可以是单个变量或 p 个变量，即

$$\boldsymbol{X}_k=(x_{1k}, x_{2k}, \cdots, x_{pk}) \qquad (k=1, 2, \cdots, n)$$

参数 $\boldsymbol{B}$ 也可以是单个参数或 m 个参数，即

$$\boldsymbol{B}=(b_1, b_2, \cdots, b_m)$$

在其后的讨论中，我们一般都假设是多个变量和多个参数的情况。有时为突出这一点，式(4.2.1)与(4.2.2)也分别写成

$$y=f(x_1, x_2, \cdots, x_p, b_1, b_2, \cdots, b_m)$$

和

$$(x_{1k}, x_{2k}, \cdots, x_{pk}, y_k) \qquad (k=1, 2, \cdots, n)$$

为简便起见，今后将用 b_i 或 b_j 表示 $b_1, b_2, \cdots, b_m$ 中的任一个。对于自变量，x_i 或 x_j 亦有类似的意义。

为使 Q 达到极小，各 b_i 应满足如下方程组：

$$\begin{cases} \dfrac{\partial Q}{\partial b_1}=0 \\ \dfrac{\partial Q}{\partial b_2}=0 \\ \quad\vdots \\ \dfrac{\partial Q}{\partial b_m}=0 \end{cases} \tag{4.2.4}$$

即残差平方和对各参数的偏导数全为零。

至此还没有对函数 $y=f(\boldsymbol{X}, \boldsymbol{B})$ 作更多的限制。如果 f 对于各 b_i 的依赖关系过于复杂，则(4.2.4)式亦将很复杂。这为 b_i 的确定带来极大的麻烦(见§4.3)。在本节，将从 y 是各 b_i 的线性函数这一简单形式入手，这时假设

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m \tag{4.2.5}$$

这里有一个参数 b_0 是实际计算的需要。为使其后讨论方便，形式上加设一个变量 $x_0\equiv1$，得

$$y=b_0x_0+b_1x_1+\cdots+b_mx_m \tag{4.2.6}$$

这时

$$Q=\sum_{k=1}^{n}[y_k-f(\boldsymbol{X}_k, \boldsymbol{B})]^2=\sum_{k=1}^{n}[y_k-(b_0x_{0k}+b_1x_{1k}+\cdots+b_mx_{mk})]^2 \tag{4.2.7}$$

$$\frac{\partial Q}{\partial b_i}=2\sum_{k=1}^{n}[y_k-(b_0x_{0k}+b_1x_{1k}+\cdots+b_mx_{mk})](-x_{ik})$$

$$=2\left(b_0\sum_{k=1}^{n}x_{0k}x_{ik}+b_1\sum_{k=1}^{n}x_{1k}x_{ik}+\cdots+b_m\sum_{k=1}^{n}x_{mk}x_{ik}-\sum_{k=1}^{n}x_{ik}y_k\right)$$

或

$$\frac{1}{2}\frac{\partial Q}{\partial b_i}=s_{i0}b_0+s_{i1}b_1+\cdots+s_{im}b_m-s_{iy} \qquad (i=0, 1, 2, \cdots, m)$$

这里

$$\begin{cases} s_{ij}=\sum\limits_{k=1}^{n}x_{ik}x_{jk} & (i, j=0, 1, 2, \cdots, m) \\ s_{iy}=\sum\limits_{k=1}^{n}x_{ik}y_k & (i=0, 1, 2, \cdots, m) \end{cases} \tag{4.2.8}$$

当观测值 $(x_{1k}, x_{2k}, \cdots, x_{mk}, y_k)$ 给出后，s_{ij} 和 s_{iy} 都可算出。这时，方程组(4.2.4)可化为

$$\begin{cases} s_{00}b_0+s_{01}b_1+\cdots+s_{0m}b_m=s_{0y} \\ s_{10}b_0+s_{11}b_1+\cdots+s_{1m}b_m=s_{1y} \\ \cdots\cdots \\ s_{m0}b_0+s_{m1}b_1+\cdots+s_{mm}b_m=s_{my} \end{cases} \tag{4.2.9}$$

这一方程组常称为“法方程”或“正规方程”，它是 $b_0, b_1, \cdots, b_m$ 的 $m+1$ 元联立方程式(即 $m+1$ 阶线代数方程组)。在 $n>m$ 的情况下，正规方程一般有唯一解。由观测数据算出系数 s_{ij} 及右端 s_{iy} 之后，便可用诸如消去法等经典方法解出 $b_0, b_1, \cdots, b_m$。在计算过程中，如果注意到系数的对称性 $s_{ij}\equiv s_{ji}$，计算量可以大为节省。

例：§4.1 中最前面的例子。

把 $y=a+bx$ 改写为

$$y=b_0+b_1x_1 \quad 或 \quad y=b_0x_0+b_1x_1$$

这里 b_0、b_1 即原来的 a、b；$x_0\equiv1$，x_1 即原来的 x。由(4.2.8)得

$$s_{00}=\sum_{k=1}^{5}x_{0k}x_{0k}=5$$

$$s_{01}=s_{10}=\sum_{k=1}^{5}x_{0k}x_{1k}=\sum_{k=1}^{5}x_{1k}=28$$

$$s_{11}=\sum_{k=1}^{5}x_{1k}x_{1k}=190$$

$$s_{0y}=\sum_{k=1}^{5}x_{0k}y_k=\sum_{k=1}^{5}y_k=18.98$$

$$s_{1y}=\sum_{k=1}^{5}x_{1k}y_k=122.83$$

由(4.2.9)得

$$\begin{cases}5b_0+28b_1=18.98\\28b_0+190b_1=122.83\end{cases}$$

解之得

$$\begin{cases}b_0=1.005783 & (\text{即 } a)\\b_1=0.498253 & (\text{即 } b)\end{cases}$$

如果对计算结果的精度有较高的要求，计算时可以把观测数据“标准化”，即作尺度变换，使变换后的各个量(记为 x_i', y')的平均值为 0，内积为 1，即

$$\frac{1}{n}\sum_{k=1}^{n}x_{ik}'=0,\quad \sum_{k=1}^{n}(x_{ik}')^2=1 \tag{4.2.10}$$

为此可令

$$x_{ik}'=\frac{x_{ik}-\bar{x}_i}{\sigma_i},\quad y_k'=\frac{y_k-\bar{y}}{\sigma_y}\quad (i=1,\ 2,\ \cdots,\ m,\quad k=1,\ 2,\ \cdots,\ n) \tag{4.2.11}$$

其中

$$\begin{cases}\bar{x}_i=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}x_{ik},\quad \sigma_i=\sqrt{\displaystyle\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}\quad (i=1,\ 2,\ \cdots,\ m)\\ \bar{y}=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}y_k,\quad \sigma_y=\sqrt{\displaystyle\sum_{k=1}^{n}(y_k-\bar{y})^2}\end{cases} \tag{4.2.12}$$

原表达式

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m$$

变换后可化为

$$y'=b_1'x_1'+b_2'x_2'+\cdots+b_m'x_m' \tag{4.2.13}$$

其中

$$b_i=b_i'\frac{\sigma_y}{\sigma_i}\qquad (i=1,\ 2,\ \cdots,\ m) \tag{4.2.14}$$

$$b_0=\bar{y}-(b_1\bar{x}_1+b_2\bar{x}_2+\cdots+b_m\bar{x}_m)^{\ominus} \tag{4.2.15}$$

由于变换后的式子(4.2.13)不显含 b_0，因此相应的正规方程也降低了一阶，得

$$\begin{cases}r_{11}b_1'+r_{12}b_2'+\cdots+r_{1m}b_m'=r_{1y}\\r_{21}b_1'+r_{22}b_2'+\cdots+r_{2m}b_m'=r_{2y}\\\cdots\cdots\\r_{m1}b_1'+r_{m2}b_2'+\cdots+r_{mm}b_m'=r_{my}\end{cases} \tag{4.2.16}$$

⊖ 其证明如下：

原表达式可写为

$$y_k=b_0+b_1x_{1k}+\cdots+b_mx_{mk}$$

$$y_k-\bar{y}=b_1(x_{1k}-\bar{x}_1)+\cdots+b_m(x_{mk}-\bar{x}_m)+b_0-\bar{y}+b_1\bar{x}_1+\cdots+b_m\bar{x}_m$$

$$\frac{y_k-\bar{y}}{\sigma_y}=b_1\frac{\sigma_1}{\sigma_y}\left(\frac{x_{1k}-\bar{x}_1}{\sigma_1}\right)+\cdots+b_m\frac{\sigma_m}{\sigma_y}\left(\frac{x_{mk}-\bar{x}_m}{\sigma_m}\right)+(b_0-\bar{y}+b_1\bar{x}_1+\cdots+b_m\bar{x}_m)/\sigma_y$$

$$y_k'=b_1'x_{1k}'+\cdots+b_m'x_{mk}'+(b_0-\bar{y}+b_1\bar{x}_1+\cdots+b_m\bar{x}_m)/\sigma_y$$

上式两边对 k 求和，并注意到

$$\sum_{k=1}^{n}x_{ik}'=\sum_{k=1}^{n}y_k'=0$$

便得(4.2.13)～(4.2.15)。

其中 r_{ij} 类似于 s_{ij}

$$\begin{cases} r_{ij}=\sum_{k=1}^{n} x'_{ik}x'_{jk}=\frac{1}{\sigma_i\sigma_j}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j) & (i,\ j=1,\ 2,\ \cdots,\ m) \\ r_{iy}=\sum_{k=1}^{n} x'_{ik}y'_{k}=\frac{1}{\sigma_i\sigma_y}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(y_k-\bar{y}) & (i=1,\ 2,\ \cdots,\ m) \end{cases} \tag{4.2.17}$$

最后指出，线性模型中参数的确定问题就是线性矛盾方程组的求解问题。因此相应章节(见第八章)的算法完全可以用在这里。

4.2.2 线性模型的推广

上面考虑的表达式

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m$$

是标准的线性式，形式简单。有一些看来较复杂的问题，往往可以通过“变量变换”的办法简化成这样的简单形式。下面可用具体例子作说明。

例 1：$y=b_0+b_1x+b_2x^2+b_3x^3$

作变量变换

$$z_1=x,\quad z_2=x^2,\quad z_3=x^3 ㊀$$

则表达式变为

$$y=b_0+b_1z_1+b_2z_2+b_3z_3$$

若把 z_1, z_2, z_3 也看作是自变量，则上式与标准的线性式无异。4.2.1 节中关于解法方程确定 b_i 的办法依然有效。当然有关式子中出现 x_i 的地方要相应地改为 z_i。

例 2：$y=a_0+\sum_{i=1}^{2}(a_i\cos ix+b_i\sin ix)$

式中的 a_0, a_1, a_2 也是待定参数，现分别改为 b_0, b_3, b_4，并作变量变换

$$z_1=\sin 1x,\quad z_2=\sin 2x$$

$$z_3=\cos 1x,\quad z_4=\cos 2x$$

得

$$y=b_0+b_1z_1+b_2z_2+b_3z_3+b_4z_4$$

这仍是标准的线性式。

综合这两个例子，可得如下的一般方法：如果原表达式为

$$y=b_0+b_1g_1(\boldsymbol{X})+b_2g_2(\boldsymbol{X})+\cdots+b_mg_m(\boldsymbol{X}) \tag{4.2.18}$$

这里的 $g_i(\boldsymbol{X})\ (i=1,\ 2,\ \cdots,\ m)$ 是 $\boldsymbol{X}=(x_1,\ x_2,\ \cdots,\ x_p)$ 的已知函数，则可作变量变换

$$z_i=g_i(\boldsymbol{X})\qquad (i=1,\ 2,\ \cdots,\ m) \tag{4.2.19}$$

把原表达式化为标准的线性式

$$y=b_0+b_1z_1+b_2z_2+\cdots+b_mz_m \tag{4.2.20}$$

使用 4.2.1 节的算法便可得到 b_i 的估值。显然原来变量的个数 p 可以不等于 m。

变量变换的方法，不仅适用于自变量，也适用于因变量 y 和待定参数 b_i。

例 3：$y=ae^{b_1x_1+b_2x_2+b_3(x_1x_2-1)}$

式中的 a 也是待定参数。今对两边取对数得

$$\ln y=\ln a+b_1x_1+b_2x_2+b_3(x_1x_2-1)$$

㊀ 写得具体一点即：$z_{1k}=x_k$, $z_{2k}=x_k^2$, $z_{3k}=x_k^3$ $(k=1,\ 2,\ \cdots,\ n)$

作变换
$$z=\ln y,\quad b_0=\ln a,\quad x_3=x_1x_2-1$$
得
$$z=b_0+b_1x_1+b_2x_2+b_3x_3$$
这是以 z 为因变量的标准线性式，可用 4.2.1 节中的方法确定 b_i；不过在得到 b_0 之后，需用逆变换式求 a，即 $a=e^{b_0}$。

这样，用于确定标准线性式参数的方法可以推广到下述形式之中：
$$h(y)=c_0(b_0)+c_1(b_1)g_1(\boldsymbol{X})+\cdots+c_m(b_m)g_m(\boldsymbol{X}) \tag{4.2.21}$$

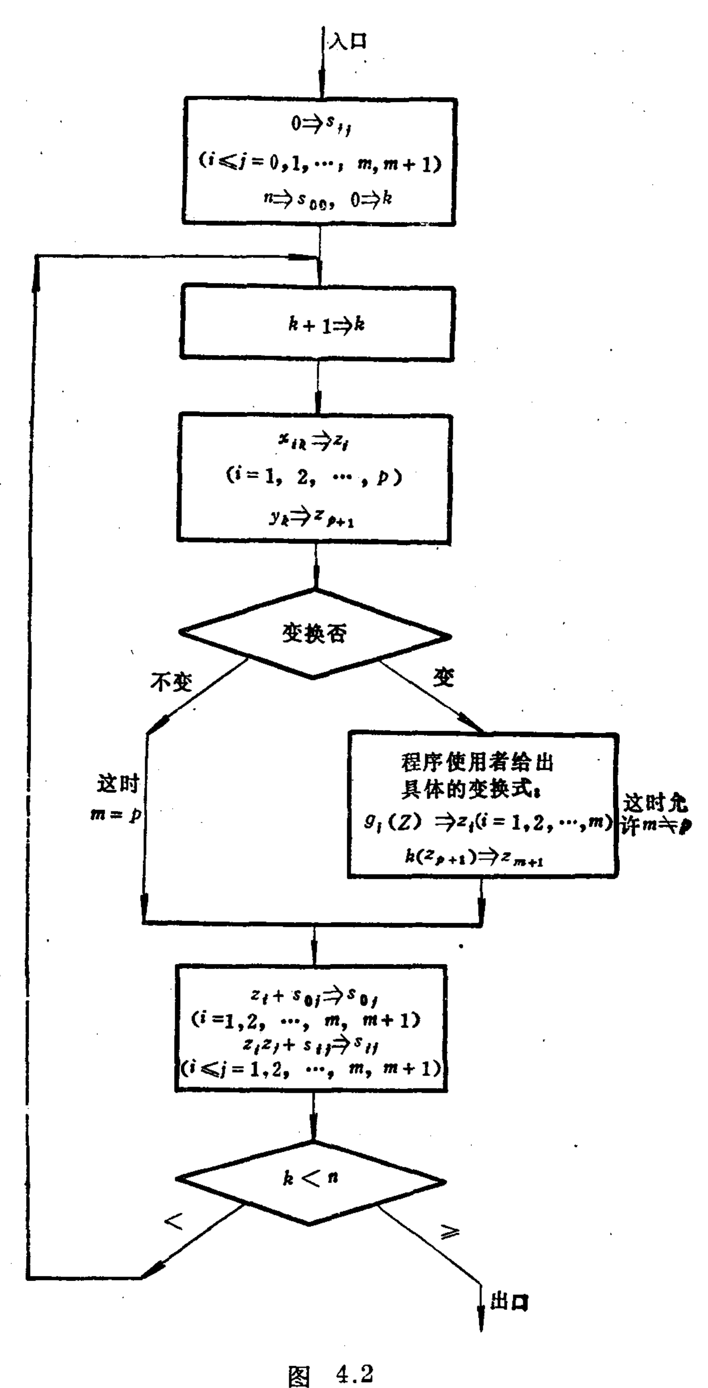


图 4.2

式中　　$h(y)$是因变量y的已知函数;

$g_i(\boldsymbol{X})$ $(i=1, 2, \cdots, m)$ 是变量 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)$的已知函数;

$c_i(b_i)$ $(i=0, 1, 2, \cdots, m)$是第 i 个参数 b_i 的已知函数。显然它必须是可逆的，以便由 $c_i(b_i)$经逆变换求得 b_i。

关于变量变换的问题,还应指出:当因变量 y 变为$h(y)$后,用 4.2.1 节的方法求出的 b_i,一般而言，已不是关于 y 的残差的最小二乘解，而只是新因变量 $h(y)$的残差的最小二乘解(在例 3 中，残差是 $\ln y-\ln y^*$)。这二者之间有差别，但是在实际计算中，人们往往只关心变量变换后可以把问题简化，而不理会这种变换可能引起解 b_i 的差异。因此，可以把式(4.2.5)与(4.2.21)都称为线性模型而不加区别。

在实际问题中，有时由于各次观测数据的精度不同以及各次观测值的差异可以超过几个数量级等原因,一般的残差平方和最小的原则已不适应,如果改用"加权"的残差平方和最小的原则更好。这时

$$Q=\sum_{k=1}^{n} W_k[y_k-(b_0+b_1x_{1k}+b_2x_{2k}+\cdots+b_mx_{mk})]^2 \tag{4.2.22}$$

这里的 $W_k(k=1, 2, \cdots, n)$是"权重",将根据实际情况给出。一般原则是:某一观测值的重要性越大,则其权重也越大。对于加权的情况,可作类似式(4.2.6～4.2.9)的推导。为确定 b_i, 仍需解(4.2.9)形状的法方程，但 s_{ij} 和 s_{iy} 要相应地改为

和
$$\begin{cases} s_{ij}=\sum_{k=1}^{n} W_k x_{ik} x_{jk} \\ s_{iy}=\sum_{k=1}^{n} W_k x_{ik} y_k \end{cases} \tag{4.2.23}$$

因变量 y 的变换,原则上可以理解为某种意义下的加权。因此,可用加权的观点来理解上面指出过的这种变换所引起的解 b_i 的变化。

最后指出，编制线性模型的标准程序时，应考虑变量变换的灵活性，可采用图 4.2 所示的框图计算 s_{ij} 及 s_{iy}(即 $s_{i,m+1}$)。

§4.3　非线性模型中参数的确定

本节要讨论的模型是最一般的模型:

$$y=f(\boldsymbol{X}, b_1, b_2, \cdots, b_m) \tag{4.3.1}$$

其中 f 可以是参数 b_i 的最一般形式的非线性函数; $\boldsymbol{X}$ 可以是单个变量，也可以是 p 个变量，即 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)$。一般的非线性问题无法直接求解。数值计算中，通常是用逐次逼近的方法处理。这种处理的实质是逐次"线性化"。

4.3.1　基本算法——高斯-牛顿法

先给 b_i 一个初始近似值,记为 $b_i^{(0)}$,并记初值与真值之差(未知的)为 Δ_i,

$$b_i=b_i^{(0)}+\Delta_i \qquad (i=1, 2, \cdots, m) \tag{4.3.2}$$

这时确定 b_i 的问题化为确定修正值 Δ_i。为确定 Δ_i, 可对函数 f 在 $b_i^{(0)}$ 附近作台劳展开，并略去 Δ_i 的二次及二次以上的项得

$$f(\boldsymbol{X}_k, b_1, b_2, \cdots, b_m)\approx f_{k0}+\frac{\partial f_{k0}}{\partial b_1}\Delta_1+\frac{\partial f_{k0}}{\partial b_2}\Delta_2+\cdots+\frac{\partial f_{k0}}{\partial b_m}\Delta_m \tag{4.3.3}$$

式中

$$\left\{\begin{aligned} & f_{k0}=f(\boldsymbol{X}_k,\ b_1^{(0)},\ b_2^{(0)},\cdots,\ b_m^{(0)}) \\ & \frac{\partial f_{k0}}{\partial b_i}=\frac{\partial f(\boldsymbol{X},\ b_1,\ b_2,\ \cdots,\ b_m)}{\partial b_i}\Bigg|_{\substack{\boldsymbol{X}=\boldsymbol{X}_k \\ b_1=b_1^{(0)} \\ \vdots \\ b_m=b_m^{(0)}}} \end{aligned}\right. \tag{4.3.4}$$

当 $b_i^{(0)}$ 给定时，它们都是自变量 $\boldsymbol{X}$ 的函数，可以直接算出。类似上节(4.2.7)的推导，并注意近似式(4.3.3)可得

$$Q=\sum_{k=1}^{n}[y_k-f(\boldsymbol{X}_k,\ b_1,\ b_2,\ \cdots,\ b_m)]^2\approx\sum_{k=1}^{n}\left[y_k-\left(f_{k0}+\frac{\partial f_{k0}}{\partial b_1}\varDelta_1+\cdots+\frac{\partial f_{k0}}{\partial b_m}\varDelta_m\right)\right]^2$$

$$\begin{aligned} \frac{\partial Q}{\partial b_i}=\frac{\partial Q}{\partial \varDelta_i}&\approx 2\sum_{k=1}^{n}\left[y_k-\left(f_{k0}+\frac{\partial f_{k0}}{\partial b_1}\varDelta_1+\cdots+\frac{\partial f_{k0}}{\partial b_m}\varDelta_m\right)\right]\left(-\frac{\partial f_{k0}}{\partial b_i}\right) \\ &=2\left[\varDelta_1\sum_{k=1}^{n}\frac{\partial f_{k0}}{\partial b_1}\frac{\partial f_{k0}}{\partial b_i}+\cdots+\varDelta_m\sum_{k=1}^{n}\frac{\partial f_{k0}}{\partial b_m}\frac{\partial f_{k0}}{\partial b_i}-\sum_{k=1}^{n}\frac{\partial f_{k0}}{\partial b_i}(y_k-f_{k0})\right] \end{aligned} \tag{4.3.5}$$

记

$$\left\{\begin{aligned} & a_{ij}=\sum_{k=1}^{n}\frac{\partial f_{k0}}{\partial b_i}\frac{\partial f_{k0}}{\partial b_j}\quad (i,\ j=1,\ 2,\ \cdots,\ m) \\ & a_{iy}=\sum_{k=1}^{n}\frac{\partial f_{k0}}{\partial b_i}\,\mathbf{X}\,(y_k-f_{k0}) \end{aligned}\right. \tag{4.3.6}$$

仍可得类似于(4.2.9)的线性方程组：

$$\left\{\begin{aligned} & a_{11}\varDelta_1+a_{12}\varDelta_2+\cdots+a_{1m}\varDelta_m=a_{1y} \\ & a_{21}\varDelta_1+a_{22}\varDelta_2+\cdots+a_{2m}\varDelta_m=a_{2y} \\ & \cdots\cdots \\ & a_{m1}\varDelta_1+a_{m2}\varDelta_2+\cdots+a_{mm}\varDelta_m=a_{my} \end{aligned}\right. \tag{4.3.7}$$

当观测值($\boldsymbol{X}_k$, y_k)给定，并给出近似值 $b_i^{(0)}$ 后，系数 a_{ij}(也是对称的)及右端 a_{iy} 可按(4.3.6)一一算出。因此可解出 $\varDelta_i$，进而得 b_i 值：

$$b_i=b_i^{(0)}+\varDelta_i\qquad (i=1,\ 2,\ \cdots,\ m) \tag{4.3.8}$$

当 $|\varDelta_i|$ 值较大时，可令当前的 b_i 值代替原来的近似值 $b_i^{(0)}$，重复(4.3.6)算出 a_{ij} 及 a_{iy} 并解(4.3.7)得新的 $\varDelta_i$ (进而得 b_i)。这种过程可以反复迭代，直至 $|\varDelta_i|$ 的值小到可以忽略为止。这时最后得到的 b_i 即为所求。这种算法常称为高斯法，也称为高斯-牛顿法或台劳展开法。其要点归纳如下：

(1) 人为地指定初值 $b_i^{(0)}$，并记 $b_i=b_i^{(0)}+\varDelta_i$，把求解 b_i 的问题化为求解 $\varDelta_i$ 的问题。

(2) 利用台劳展开式，$\varDelta_i$ 即为线性模型的待定参数，借用上节方法可求出 $\varDelta_i$(即把这里的 y_k-f_{k0} 看作是上节的因变量 y_k，$\dfrac{\partial f_{k0}}{\partial b_i}$ 看作是自变量 x_{ik})。这一过程亦称线性化。

(3) 以 $\varDelta_i$ 修正 $b_i^{(0)}$ 作为新的初值，重复(1)、(2)直至 $|\varDelta_i|$ 小于允许误差，再次修正已成为不必要。

求解过程之所以需要反复迭代和修正，是因为关键的台劳展开式(4.3.3)只是近似的式子，因此得到的 b_i 也是近似的，其近似程度依赖于 $|\varDelta_i|$ 的大小。若 $|\varDelta_i|$ 较大，$\varDelta_i$ 的二次及二次以上的项就不能忽略，台劳展式的近似性就较差。但一般而言，经一次修正后得到的

b_i 虽然还不是真正的解，却可能比原来的 $b_i^{(0)}$ 更接近于真解；逐次迭代的结果，将使当前的 b_i 值逐步逼近真解。这也可以理解为修正量 $|\Delta_i|$ 逐次缩小。当 $|\Delta_i|$ 小到它的二次项可以忽略时，台劳展开式(4.3.3)事实上可以看作是精确的式子㊀，这便得到真解 b_i。如果迭代过程果真按照上述方式完成，则迭代称为"收敛的"。显然，在这过程中可能包含着大量的反复计算。

非线性问题的难点，远不是很大的计算量，而是迭代过程有可能不按上述方式完成，即所谓"发散"现象。这是因为初值 $b_i^{(0)}$ 选得不好，台劳展开式(4.3.3)完全失真，迭代得到的新 b_i 有可能比原来的 $b_i^{(0)}$ 更远离真解，而且越迭代越糟糕。迭代的收敛或发散的关键在于初值 $b_i^{(0)}$ 的选择。对于不少问题，要选一个较好的初值有时又是困难的，为了放宽对初值的限制，麦夸脱提出了修改方案[1]。

4.3.2 算法改进——麦夸脱法

麦夸脱法与高斯-牛顿法相似，其差别仅在于迭代时确定 Δ_i 的线性代数方程组改为：

$$\begin{cases}(a_{11}+d)\Delta_1+a_{12}\Delta_2+\cdots+a_{1m}\Delta_m=a_{1y}\\ a_{21}\Delta_1+(a_{22}+d)\Delta_2+\cdots+a_{2m}\Delta_m=a_{2y}\\ \cdots\cdots\\ a_{m1}\Delta_1+a_{m2}\Delta_2+\cdots+(a_{mm}+d)\Delta_m=a_{my}\end{cases} \tag{4.3.9}$$

即在原方程(4.3.7)的对角线上加了一个因子 d $(d\geqslant 0)$㊁。显然，当 $d=0$ 时麦夸脱法退化为高斯-牛顿法。若引入记号

$$\boldsymbol{\Delta}=(\Delta_1,\ \Delta_2,\ \cdots,\ \Delta_m)$$
$$\boldsymbol{a}_y=(a_{1y},\ a_{2y},\ \cdots,\ a_{my})$$

可以证明(参看[1])：

(1) 当 d 越来越大时，$\boldsymbol{\Delta}$ 的长度越来越小，并以零为极限，即

$$\lim_{d\to\infty}\|\boldsymbol{\Delta}\|=0 \tag{4.3.10}$$

(2) 当 d 越来越大时，$\boldsymbol{\Delta}$ 和 $\boldsymbol{a}_y$ 两矢量的夹角 γ 越来越小，并以零为极限，即

$$\lim_{d\to\infty}\gamma=0 \tag{4.3.11}$$

由于 $\boldsymbol{a}_y$ 不随 d 改变，(2)的结论实际上指出 $\boldsymbol{\Delta}$ 的方向将随着 d 的增大而逐渐接近 $\boldsymbol{a}_y$ 的方向。又由于 $\boldsymbol{a}_y$ 方向就是梯度方向(最速下降方向)，沿着这个方向，只要步长不太大，残差平方和 Q 总可以逐渐减少。所以，只要 d 充分大，定能保证下一次迭代中得到的 Q 值比上一次的小，除非当前的 $b_i^{(0)}$ 值已是所求的真解了。

㊀ 被忽略的 Δ_i 的二次及二次以上的项的系数依赖于二阶及二阶以上的偏导数。如果这些偏导数全为零，则不论 Δ_i 有多大，式(4.3.3)总是精确成立的。因此 b_i 不需再迭代。§4.2中的线性模型正是这种情况。在那里特别令 $b_i^{(0)}=0$，则有 $f_{k0}=0$，$\frac{\partial f_{k0}}{\partial b_i}=x_{ik}$。因而 $a_{ij}=s_{ij}$，$a_{iy}=s_{iy}$，并且解出的 Δ_i 就是 b_i 的真解。然而在非线性模型中，这些偏导数在解的附近不可能全为零，我们只能假设它们都存在并且有界。

㊁ 式(4.3.9)实际上是使

$$Q'=\sum_{k=1}^{n}\left[y_k-\left(f_{k0}+\frac{\partial f_{k0}}{\partial b_1}\Delta_1+\cdots+\frac{\partial f_{k0}}{\partial b_m}\Delta_m\right)\right]^2+d\sum_{i=1}^{m}\Delta_i^2$$

达到最小的 Δ_i 应满足的方程，这时 d 亦称为"阻尼"因子。

上面的两个结论，提供了因子 d 的选取原则，亦即在收敛的情况下，为减少迭代次数，d 宜选取较小的值；仅当不能保证相应的 Q 值比前次小的情况下，才选较大的 d 值。因此，d 是随迭代过程变化的。例如可以给出如下的具体选取方法：

(1) 算出初值 $b_i^{(0)}$ 所对应的残差平方和，并记为 $Q^{(0)}$；指定一个常数 c $(c>1$，例如令 $c=10)$；并给出 d 的一个初值 $d^{(0)}$（例如当 $a_{ii}=1$ 时，令 $d^{(0)}=0.01$）。

(2) 进行下一次迭代时，令

$$d=c^{\alpha}d^{(0)} \qquad (\alpha=-1,\ 0,\ 1,\ 2,\ 3,\ \cdots) \tag{4.3.12}$$

这里的 α 值尽可能取得小些，但需保证式(4.3.9)解出的 Δ_i（进而是 b_i）相应的残差平方和 Q 不大于 $Q^{(0)}$，即

$$Q<Q^{(0)} \tag{4.3.13}$$

也就是说，先令 $d=c^{-1}d^{(0)}=d^{(0)}/c$，若(4.3.13)成立，则这一次迭代完成；否则令 $d=c^0d^{(0)}=d^{(0)}$，若(4.3.13)成立，则这一次迭代完成；否则令 $d=c^1d^{(0)}$，…。根据上面的论证，只要 $d=c^{\alpha}d^{(0)}$ 充分大，总能保证(4.3.13)成立，从而结束这次迭代。

(3) 以 d，b_i 和 Q 的当前值（即结束上一次迭代时的值）代替 $d^{(0)}$，$b_i^{(0)}$ 和 $Q^{(0)}$，重复(2)作下一次迭代，直至 $|\Delta_i|$ 可以忽略，迭代过程收敛㊀。

这便是麦夸脱方法的主要思想。具体的计算程序将在4.3.4节中给出。计算经验表明，这一方法确实比高斯-牛顿法有效。4.3.3节中的实例计算结果表明了这一点。在实际计算中，高斯-牛顿法或麦夸脱法都可以进一步改进。例如当偏导数太复杂时，可用差商代替[2,3]；又如为避免在寻求合适的 d 的过程中反复求解线性方程组(4.3.9)，可用近似的 Δ 来替代[4]，甚至可用其它方法选取合适的 d 值[5]等等。

4.3.3 实例与算法比较

晶体振荡器的频率是随温度变化的。为使频率稳定，简单的办法是使用如图4.3所示的热敏网络来补偿㊁。网络的输出电压 V 随着温度 T 变化。它作用于振荡器之后，能使振荡器频率也随 T 变化，但要求这种变化恰好与原来的变化相反，以便相互抵消，使频率最终得以稳定。

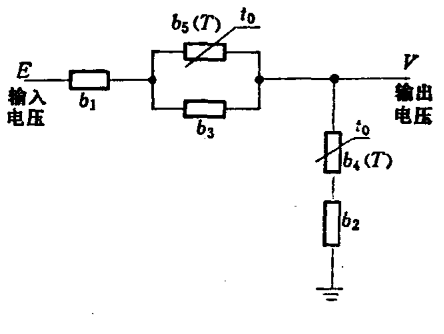


图4.3 热敏网络

我们的问题是：经测试，知某振荡器需按图4.4（或表4.2）所示的温度-电压关系来补偿，要求确定补偿网络中各参数的值（包括电阻 a_i，b_i 及变化参数 c_i）。

㊀ 有时虽然尚未收敛，但由于 $d\to\infty$，也可以引起 $|\Delta_i|\to 0$，因此需仔细鉴别，见4.3.4节的程序。

㊁ 图4.3中，b_i 表示阻值为 b_i 的电阻；

$b_i(T)t_0$ 表示阻值随温度变化的热敏电阻

$$b_i(T)=a_ie^{c_i\left(\frac{1}{T}-\frac{1}{293}\right)}$$

这里 c_i 为变化参数；

a_i 为293°K（即20°C）时本电阻之阻值；

温度 T 用绝对温度表示。

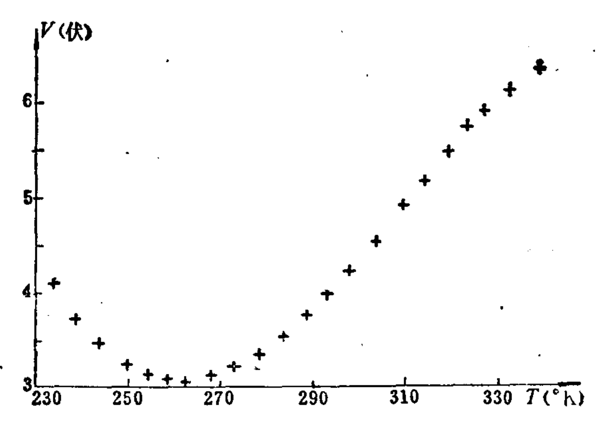


图 4.4 温度-电压曲线

表 4.2 (T_k, V_k) 数据

k	T_k °K	V_k 伏	k	T_k °K	V_k 伏
1	233.4025	4.202	12	288.66575	3.750
2	239.35995	3.730	13	292.99200	3.966
3	243.59725	3.493	14	297.68075	4.210
4	249.90110	3.252	15	303.37210	4.527
5	254.20285	3.155	16	309.73835	4.912
6	258.05660	3.110	17	314.01480	5.168
7	262.54575	3.086	18	319.24615	5.469
8	268.73452	3.151	19	323.76605	5.713
9	272.85824	3.218	20	327.75165	5.905
10	278.15304	3.357	21	332.77260	6.122
11	283.27835	3.530	22	339.09160	6.356

注：温度 T 用绝对温度表示，温度的读数是换算出来的，因此数位较多。

由欧姆定律，网络中各量的关系如下：

$$\frac{E-V}{V}=\frac{b_1+\dfrac{b_3\cdot a_5e^{c_5\left(\frac{1}{T}-\frac{1}{293}\right)}}{b_3+a_5e^{c_5\left(\frac{1}{T}-\frac{1}{293}\right)}}}{b_2+a_4e^{c_4\left(\frac{1}{T}-\frac{1}{293}\right)}} \tag{4.3.14}$$

$$V=\frac{E\cdot\left(b_2+a_4e^{c_4\left(\frac{1}{T}-\frac{1}{293}\right)}\right)}{b_1+\dfrac{b_3\cdot a_5e^{c_5\left(\frac{1}{T}-\frac{1}{293}\right)}}{b_3+a_5e^{c_5\left(\frac{1}{T}-\frac{1}{293}\right)}}+b_2+a_4e^{c_4\left(\frac{1}{T}-\frac{1}{293}\right)}} \tag{4.3.15}$$

式中输入电压 $E=12$ 伏，是已知常数；$a_5=1$ 也是已知常数（这是假设的，它将使解有唯一性，并使计算简化。如果实际使用时，$a_5=100$，则 b_1, b_2, b_3 和 a_4 都应乘 100 倍）。因此网络中各参数的确定问题，就是曲线拟合中的参数确定问题，其表达式是(4.3.15)，待定参数是 $b_1, b_2, b_3, a_4, c_4, c_5 (m=6)$，观测数据是表 4.2 中的 (T_k, V_k) $k=1, 2, \cdots, n (n=22)$。

对不同的初值，使用高斯-牛顿法和麦夸脱法的计算情况见表 4.3 至表 4.6。

表 4.3　高斯-牛顿法对几个初值的计算情况

初　值　$a_4^{(0)}$	0.02	0.03	0.04	0.05	0.1	0.2
初值对应的残差平方和 $Q^{(0)}$	6.8472	13.5382	21.8361	31.0192	79.9855	166.9445
收敛时的迭代次数	7	7	8	失败	失败	失败

表 4.4　麦夸脱法对几个初值的计算情况

初　值　$a_4^{(0)}$	0.02	0.1	0.4	0.5	0.6	1.0
初值对应的残差平方和 $Q^{(0)}$	6.8472	79.9855	293.1680	341.1420	382.5730	507.2890
收敛时的迭代次数	6	18	9	失败	失败	失败

表 4.5　两种方法迭代过程对比 $a_4^{(0)}=0.02,\ Q^{(0)}=6.8472$

迭代次数		1	2	3	4	5	6	7
高斯-牛顿法	b_1	−0.02205450	0.06552070	0.1758245	0.2133135	0.2147570	0.2147405	0.2147405
	b_2	0.2218455	0.2746170	0.3381135	0.3558130	0.3569820	0.3569855	0.3569855
	b_3	1.084085	1.100250	1.164690	1.135845	1.141900	1.142065	1.142070
	a_4	0.010062	0.01306475	0.01328110	0.01141960	0.01161490	0.01162095	0.01162090
	c_4	4258.000	4034.290	4473.825	4695.850	4698.505	4698.075	4698.075
	c_5	3737.605	3674.360	3824.090	3951.035	3950.975	3950.680	3950.680
	Q	9.979615	0.7983040	0.03474510	0.00121607	0.00091597	0.00091595	0.00091595
麦夸脱法	b_1	0.1449170	0.1996885	0.2141980	0.2147445	0.2147405	0.2147405	程序控制当 $d=0.000001$ 时不再减小
	b_2	0.3252305	0.3483100	0.3565255	0.3569815	0.3569855	0.3569855	
	b_3	1.218880	1.122550	1.139300	1.141985	1.142065	1.142065	
	a_4	0.01576230	0.01121155	0.01152395	0.01161775	0.01162085	0.01162090	
	c_4	5000.965	4716.095	4701.390	4698.240	4698.080	4698.075	
	c_5	3645.965	3904.435	3957.280	3950.855	3950.680	3950.680	
	Q	1.533985	0.03188835	0.00094167	0.00091596	0.00091595	0.00091595	
	d	0.001	0.0001	0.00001	0.000001	0.000001	0.000001	

表 4.6　初值稍差时麦夸脱法迭代过程中 d 的变化 $a_4^{(0)}=0.1$

迭代次数	0	1	2	3	4	5	6
Q	79.98555	12.46410	8.753780	2.884910	0.2829345	0.03519820	0.02056185
d		0.001	0.0001	10	1	0.1	0.01
迭代次数	7	8	9	10	11	12	13
Q	0.00552958	0.00152503	0.00115947	0.00104580	0.00099057	0.00096073	0.00094245
d	0.001	0.001	0.001	0.001	0.001	0.001	0.001
迭代次数	14	15	16	17	18	程序控制当 $d=0.000001$ 时不再减小	
Q	0.00093310	0.00092870	0.00091623	0.00091595	0.00091594		
d	0.001	0.0001	0.00001	0.000001	0.000001		

计算是在 109 乙机上进行的。为便于对比，初值的 6 个参数中，只对较敏感的 a_4 在区间(0.02, 1)上取了不同的初值，其它固定不变，即

$$b_1^{(0)}=0.5,\ b_2^{(0)}=0.5,\ b_3^{(0)}=1,\ c_4^{(0)}=5000,\ c_5^{(0)}=4000$$

若 $|\Delta_i|/(|b_i|+0.001)<0.00001$，则认为迭代收敛，结果为

$$b_1=0.2147405,\ b_2=0.3569855,\ b_3=1.142070$$

$$a_4=0.01162090,\ c_4=4698.075,\ c_5=3950.680$$

残差平方和 $Q=0.00091595$

计算结果表明，对于高斯-牛顿法，$a_4^{(0)}=0.02,\ 0.03,\ 0.04$ 都可获得上述结果，但 $a_4^{(0)}=0.05$ 或更大时便告失败。若用麦夸脱法，则当 $a_4^{(0)}\leqslant 0.4$ 时都能获得上述结果，这比高斯-牛顿法对初值的要求放宽了一个数量级。

上述例子是有代表性的。计算经验表明，当初值较好时，麦夸脱法计算顺利，d 值逐步下降，收敛过程与高斯-牛顿法没有多大差别。这时对于相同的允许误差而言，两个方法的迭代次数一般相差甚少(见表 4.5)。对于高斯-牛顿法不收敛的初值，使用麦夸脱法则有可能收敛，而且有时迭代次数也不很多(见表 4.4，特别是 $a_4^{(0)}=0.4$ 时)。但完全可能出现另外的情况，即在迭代的开头几步，d 值一下子变得很大，大大影响了收敛速度(见表 4.6)。在这种情况下，我们势必为寻找结果耗费一些时间，但总比不能求得结果好一些。当然麦夸脱法放宽对初值的要求也不是没有限制的。实在太差的初值，可能在迭代计算中导致异常情况，或掉进"坑里"。计算以失败告终(见表 4.4 中 $a_4^{(0)}\geqslant 0.5$ 的情况)。在上例中，当 $a_4^{(0)}=0.5$ 时，迭代四步以后，Q 值从 341.1420 下降到 0.9920385。但此后虽经过百次迭代，所求参数还是在不合理的区域上徘徊，致使 Q 值总不小于 0.35。

4.3.4 程序

过程 LSN(M, N, B, G, EE, F, PD, E, ERR, MAXP, GN, LF);

使用说明

本过程按最小二乘意义估计经验公式中的非线性参数。方法包括麦夸脱法与高斯-牛顿法。算法与记号如前，但解线性方程组求 Δ 时(这里改为 G)，先将 a_{ij} 标准化，即用 $a'_{ij}=a_{ij}/\sqrt{a_{ii}}\sqrt{a_{jj}}$ 代替原来的 a_{ij}，若记替换后的解为 b'_i，则 $b_i=b'_i\sqrt{a_{yy}}/\sqrt{a_{ii}}$(参看式(4.2.14))。此外，当 $d\geqslant 100$ 时，不再增大，改用 $b_i=b_i^{(0)}+w\Delta_i$，其中 $0<w<1$。

现将各参数介绍如下：

使用者需定义三个场：B[1:M]，G[1:M+1]，EE[1:H]。M 为待定参数个数。N 为观测次数。B 在计算前放初值，计算后为结果。F 是自编算残差的过程。PD 是自编算偏导数的过程。说明时，这两个过程都要带参数，如过程 F(K)，过程 PD(K)，其中 K 为值，用以指示当前应计算的观测点的顺序号，计算时，参数依次从 B[1]，B[2]，…，B[M]取出，算得的残差放在 E 内，偏导数依次放在 G[1]，G[2]，…，G[M] 内。ERR 是结束迭代的允许误差(相对误差)。MAXP 是最高的迭代次数。在 GN=1 时，本过程用高斯-牛顿法，其它情况用麦夸脱法。计算结束时，各残差依次放在场 EE 内，残差平方和放在 E 内。LF 是开关，它应含有三个标号，依次处理如下三种意外情况：① 迭代次数已达 MAXP 仍未收敛；② 麦夸脱法的因子 $d=100$ 时改令 $b_i=b_i^{(0)}+w\Delta_i$，但 $w=0$ 也不能使残差平方和下降；③ 迭代中矩阵(a_{ij})的主元素为零。

熟悉非线性代数方程但求解的读者，不难看出，本过程也可用于此目的。这时把 B 作为所求的未知数，过程 F(K) 改为计算 0 减去第 K 个方程式的值(即 −1 乘该方程的值)。

```
过程  LSN(M, N, B, G, EE, F, PD, E, ERR, MAXP, GN, LF);
   值  M, N, ERR, MAXP, GN; 简变 E;
   场  B, G, EE; 过程 F, PD; 开关 LF;
始  简变 D, D0, Q, Q0, P, W;
   场  B0[1:M], H[1:M+1], A, A0[1:M, 1:M+1];
   过程  SUM;
     始  0⇒Q;
       对于 K=1 到 N 步长 1 执行
       始  F(K); E⇒EE[K]; E*E+Q⇒Q 终
     终;
   L0:0⇒P;
     若 GN=1 则 0⇒D 否 0.01⇒D; SUM;
   LP:P+1⇒P; Q⇒Q0; B⇒B0; 0⇒A0;
   LA: 对于 K=1 到 N 步长 1 执行
       始  PD(K); EE(K)⇒G[M+1];
           对于 I=1 到 M 步长 1 执行
             始  G[I]⇒Q;
                 对于 J=I 到 M+1 步长 1 执行
                     Q*G[J]+A0[I, J]⇒A0[I, J]
             终
       终;
       对于 I=1 到 M 步长 1 执行
       始  A0[I, I]⇒Q;
           若 Q≤0 则转 LF[3] 否 1/§SQRT(Q)⇒H[I]
       终;
       1/§SQRT(Q0)⇒H[M+1];
       对于 I=1 到 M 步长 1 执行
       始  H[I]⇒Q;
           对于 J=I+1 到 M+1 步长 1 执行 A0[I, J]*Q*H[J]⇒A0[I, J];
           Q/H[M+1]⇒  H[I]
       终;
   LD1: 若 D<0.0000002 则否 D/10⇒D; 1⇒W; D⇒D0;
   LB:  A0⇒A;
       对于 I=1 到 M 步长 1 执行 1+D⇒A[I, I];
       对于 I=1 到 M 步长 1 执行
       始  对于 J=I+1 到 M 步长 1 执行 A[I, J]⇒A[J, I];
           若 A[I, I]=0 则转 LF[3] 否 1/A[I, I]⇒Q;
```

```
              对于 J=I+1 到 M+1 步长 1 执行
                A[I, J]*Q⇒A[I, J];
              对于 K=I+1 到 M 步长 1 执行
                始  A[K, I]⇒Q;
                    对于 J=K 到 M+1 步长 1 执行 A[K,J]-Q*A[I,J]⇒A[K, J]
                终
          终;
          对于 I=M 到 1 步长 -1 执行
          始  对于 J=I+1 到 M 步长 1 执行
                  A[I, M+1]-A[J, M+1]*A[I, J]⇒A[I, M+1];
                  A[I, M+1]*H[I]⇒G[I]; B0[I]+G[I]⇒B[I]
          终;
  LERR:SUM;
        若 D0<D 则否
          始  对于 I=1 到 M 步长 1 执行
              若 §ABS(G[I])/(§ABS(B[I])+0.001)≤ERR 则否转 LD2;
              转 LEND
          终;
  LD2:  若 Q<Q0 则若 P<MAXP 则转 LP 否转 LF[1]否;
        若 GN=1 则
          若 P<MAXP 则转 LP 否转 LF[1] 否;
        若 D<20 则始 10*D⇒D; 转 LB 终否;
  LW:   W/4⇒W;
        若 W=0 则转 LF[2] 否;
        对于 I=1 到 M 步长 1 执行
          B0[I]+W*G[I]⇒B[I];
        SUM; 若 Q<Q0 则转 LP 否转 LW;
  LEND: Q⇒E
终
```

§4.4 借助数学方法选取表达式

4.4.1 问题的提出

前面两节，我们介绍了表达式中待定参数的确定方法，但这不是曲线拟合问题的全部。曲线拟合问题中首先需要解决的，是给出表达式的形式，其后才是确定这个表达式中的待定参数。一些简单的问题，或变量之间已有较明确物理关系的问题，要给出一个表达式并不太困难。但是较复杂的问题，要建立一个有效的表达式就不容易了。在一些问题中，可供选取的自变量很多，我们不但搞不清这些自变量及待定参数应该以什么样的形式出现在表达式中，甚至也难于辨明哪个变量或参数不重要，可以舍去。这就提出了表达式的选取与变量的

选取问题。

例如，在气象预报中，预报式的自变量可以有当地的气温、气压和湿度，以及邻近一、二个或更多个地区的气温、气压和湿度。甚至太阳黑子及大气环流方面的指标与此都有关系。而且这些指标可以是今天的数据，也可以是昨天、前天或更早的数据。这既有选取表达式的问题，也有选取重要变量的问题。

又如，已知函数 $f(x)$ 在 $x_1, x_2, \cdots, x_n$ 处的值为 $f_1, f_2, \cdots, f_n$，今用多项式作逼近

$$y=b_0+b_1x^1+b_2x^2+\cdots+b_mx^m$$

根据数学定理，m 越大，逼近的精度越高。但实际计算表明，m 过大时，不但求解过程中容易发生病态等麻烦情况，而且得到的多项式尽管在各 x_k 处的值与 f_k 很接近，但其它地方却产生不合理的波动现象。所以，在实际计算中应避免选项过多。这也提出了从 $x^1, x^2, \cdots, x^m$ 中适当地删掉一些项的问题。

更常见的例子是实验数据、曲线、图表的公式化。这是为进一步掌握实验结果内在关系的常用方法。但是，使用电子计算机作计算时，为了输入和计算的方便，有时也需要把大量的数据、曲线、图表公式化。在石油、化工的计算中，这一点较突出。

本节将介绍一种数学技巧——正交筛选法[6]。经验表明，它对选取重要变量和建立表达式是有效的。当然，由于实际问题的复杂性，要想较好地解决问题，单靠数学技巧或对各种函数图象的熟悉是不够的；必须深入实际，作一番调查研究和全面分析，搞清其物理背景，才能减少盲目性，并收到效果。在这个过程中，反复计算也许是不可避免的。

下面，首先介绍选取重要变量的方法(见 4.4.2 节)，然后把这个方法应用于选取表达式上(见 4.4.4 节)。实际上，表达式的形式仅限于式(4.4.35)所示的一类之中。

4.4.2 变量的正交筛选法

这里讨论在线性表达式中选取重要变量的办法。令表达式为

$$y=b_1x_1+b_2x_2+\cdots+b_mx_m \tag{4.4.1}$$

观测数据为

$$(x_{1k},\ x_{2k},\ \cdots,\ x_{mk},\ y_k) \qquad (k=1,\ 2,\ \cdots,\ m)$$

为其后讨论方便，这里假设⊖各个变量都是标准化了的，即平均值为 0，内积为 1：

$$\begin{cases} \bar{x}_i=\dfrac{1}{n}\sum\limits_{k=1}^{n}x_{ik}=0, & \sum\limits_{k=1}^{n}x_{ik}^2=1 \\ \bar{y}=\dfrac{1}{n}\sum\limits_{k=1}^{n}y_k=0, & \sum\limits_{k=1}^{n}y_k^2=1 \end{cases} \qquad (i=1,\ 2,\ \cdots,\ m) \tag{4.4.2}$$

问题是从 m 个自变量 $x_1, x_2, \cdots, x_m$ 中选取几个对表达 y 是重要的变量。正交筛选法正是解决这种问题的方法。

在描述筛选法之前，先引入如下的量

⊖ 这一假设不是必要的，而且不失一般性，因为由 §4.2 中式 (4.2.10)～(4.2.15) 的讨论可知，对于常见的另一形式

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m$$

可通过变量变换(标准化)化为(4.4.1)的形式。

$$r_{iy}=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(y_k-\bar{y})^2}} \tag{4.4.3}$$

r_{iy} 称为 x_i 与 y 的(线性)相关系数，$|r_{iy}|$ 的大小标志着两个量中的一个量可用另一个量线性表出的程度。例如 y 就是 x_i 的线性函数，即 $y=a\pm bx_i\,(b>0)$，则 $r_{iy}=\pm1$，其它情况 $|r_{iy}|<1$。若假设 y 是标准化的量，x_i 的平均值为零，则可简写为

$$r_{iy}=\sum_{k=1}^{n}x_{ik}y_k\Big/\sqrt{\sum_{k=1}^{n}x_{ik}^2} \tag{4.4.4}$$

若假设 y 与 x_i 都是标准化的量，则

$$r_{iy}=\sum_{k=1}^{n}x_{ik}y_k \tag{4.4.5}$$

这就是内积。

下面可以看到相关系数 r_{iy} 的作用。

一般当 y 用 x_i 线性表出时：

$$y=bx_i \tag{4.4.6}$$

其待定参数 b 由 §4.2 得

$$b=\sum_{k=1}^{n}x_{ik}y_k\Big/\sum_{k=1}^{n}x_{ik}^2 \tag{4.4.7}$$

在 x_i, y 已标准化的假设下，易知 $b=r_{iy}$，即

$$y=r_{iy}x_i \tag{4.4.8}$$

这时残差平方和

$$\begin{aligned}Q=\sum_{k=1}^{n}(y_k-r_{iy}x_{ik})^2&=\sum_{k=1}^{n}y_k^2-2r_{iy}\sum_{k=1}^{n}x_{ik}y_k+r_{iy}^2\sum_{k=1}^{n}x_{ik}^2\\&=1-2r_{iy}r_{iy}+r_{iy}^2=1-r_{iy}^2\end{aligned} \tag{4.4.9}$$

可见从 $x_1, x_2, \cdots, x_m$ 中选一个对表达 y 最好(即残差平方和 Q 最小)的变量，就是选使 $|r_{iy}|$ 最大的变量 x_i。换句话说，r_{iy} 可作为变量选取的标准。

正交筛选的具体计算步骤如下：

(1) 选第一个变量：即在 x_i 中选对表达 y 最好的变量。如上所述，只需算出各 x_i 与 y 的相关系数㊀：

$$r_{iy}=\sum_{k=1}^{n}x_{ik}y_k\Big/\sqrt{\sum_{k=1}^{n}x_{ik}^2} \tag{4.4.10}$$

并选 $|r_{iy}|$ 最大者，则相应的 x_i 即为第一个被选中的量。为方便起见，假设恰好选中 x_1，并令

$$r_1=r_{1y} \tag{4.4.11}$$

$$z_1=x_1\Big/\sqrt{\sum_{k=1}^{n}x_{1k}^2} \tag{4.4.12}$$

则得最优的单变量表达式

$$y=r_1z_1 \tag{4.4.13}$$

㊀ 由于 x_i 已标准化，$\sum_{k=1}^{n}x_{ik}^2=1$，因此分母可以不写出，这里保留分母的目的在于与其后的形式一致，其后 $\sum_{k=1}^{n}(x_{ik}^{(l)})^2\neq1$。

(2) 选第二个变量:

(i) 对未被选中的变量作变换使之与 z_1 正交, 可得 $x_i^{(1)}$,

$$x_i^{(1)}=x_i-c_{1i}z_1 \qquad (i=2, 3, \cdots, m) \tag{4.4.14}$$

这里的 c_{1i} 由 $x_i^{(1)}$ 与 z_1 正交这个条件

$$\sum_{k=1}^{n} x_{ik}^{(1)} z_{1k}=0 \tag{4.4.15}$$

来确定, 即

$$c_{1i}=\sum_{k=1}^{n} x_{ik}z_{1k} \tag{4.4.16}$$

(ii) 对 y 扣除用 z_1 线性表出的部分得 $y^{(1)}$㊀,

$$y^{(1)}=y-r_1z_1 \tag{4.4.17}$$

并在 $x_i^{(1)}$ 中选对表达 $y^{(1)}$ 最好的变量作为第二个被选中的量。原则上可用类似(1)的办法算出 $x_i^{(1)}$ 与 $y^{(1)}$ 的相关系数并选绝对值最大者。在实际计算中, 考虑到 $x_i^{(1)}$ 与 z_1 正交, 从而 $x_i^{(1)}$ 与 $y^{(1)}$ 的相关系数等于 $x_i^{(1)}$ 与 y 的相关系数, 因此可以不算出 $y^{(1)}$, 并改为算 $x_i^{(1)}$ 与 y 的相关系数, 并选绝对值最大者:

$$r_{iy}^{(1)}=\sum_{k=1}^{n} x_{ik}^{(1)}y_k \Big/ \sqrt{\sum_{k=1}^{n} (x_i^{(1)})^2} \tag{4.4.18}$$

假设 $|r_{2y}^{(1)}|$ 最大, 即 $x_2^{(1)}$ 被选中, 令

$$r_2=r_{2y}^{(1)} \tag{4.4.19}$$

$$z_2=x_2^{(1)} \Big/ \sqrt{\sum_{k=1}^{n} (x_{2k}^{(1)})^2} \tag{4.4.20}$$

则得 y 的两个变量的表达式

$$y=r_1z_1+r_2z_2 \tag{4.4.21}$$

(3) 类似的算法可选出第 3, 4, …, s 个变量。在此基础上再选第 $s+1$ 个变量的方法是:

(i) 作变量变换得新的 $x_i^{(s)}$:

$$x_i^{(s)}=x_i^{(s-1)}-c_{si}z_s \qquad (i=s+1, s+2, \cdots, m) \tag{4.4.22}$$

其中

$$c_{si}=\sum_{k=1}^{n} x_{ik}^{(s-1)}z_{sk} \tag{4.4.23}$$

以保证 $x_i^{(s)}$ 与 z_s 正交。

(ii) 计算相关系数

$$r_{iy}^{(s)}=\sum_{k=1}^{n} x_{ik}^{(s)}y_k \Big/ \sqrt{\sum_{k=1}^{n} (x_{ik}^{(s)})^2} \tag{4.4.24}$$

并选 $|r_{iy}^{(s)}|$ 的最大者, 以确定第 $s+1$ 个被选中的量。设 $x_{s+1}^{(s)}$ 被选中, 令

$$r_{s+1}=r_{s+1,y}^{(s)} \tag{4.4.25}$$

$$z_{s+1}=x_{s+1}^{(s)} \Big/ \sqrt{\sum_{k=1}^{n} (x_{s+1,k}^{(s)})^2} \tag{4.4.26}$$

则得 y 的 $s+1$ 个变量的表达式

$$y=r_1z_1+r_2z_2+\cdots+r_{s+1}z_{s+1} \tag{4.4.27}$$

㊀ $x_i^{(1)}$ 与 $y^{(1)}$ 实际上是用 x_1(即 z_1)表达 x_i 与 y 所得的残差。

(4) 重要变量个数的确定：上述过程只解决每次从指定的变量中选出一个相对其它为优的变量，它没有保证这个相对最优的量确实对表达 y 有显著的作用。因此，要给出一个“显著”的标准。例如可以指定一个 r_0 值，若

$$|r_s| \geqslant r_0 \tag{4.4.28}$$

则认为第 s 个变量作用确实“显著”，准其选入，并作下一变量的选取试验；否则不选入，并结束筛选过程，这时有 $s-1$ 个变量被选中。在实际计算中，可借助数理统计的方法，建立一个比(4.4.28)更好的检验标准，即构造一个量 F㊀：

$$F = \frac{(n-s-1)r_s^2}{1-\sum_{i=1}^{s} r_i^2} \tag{4.4.29}$$

并指定一个临界值 F_0，以下式代替(4.4.28)：

$$F \geqslant F_0 \tag{4.4.30}$$

一般取 $F_0=4$。若希望多选一些变量，则 F_0 值稍小些，例如 $F_0=2$，$F_0=1$。反之 F_0 值可稍大些，例如 $F_0=7$，$F_0=10$ 等。

(5) 从变量 z_i 转换为 x_i 的计算：上述计算过程所得的 y 的表达式是关于 z_i 的，即

$$y = r_1 z_1 + r_2 z_2 + \cdots + r_s z_s \tag{4.4.31}$$

在使用时，常希望得到关于 x_i 的表达式

$$y = b_1 x_1 + b_2 x_2 + \cdots + b_s x_s \tag{4.4.32}$$

这时可逐次使用式(4.4.26)，(4.4.22)作逆变换。必须指出，这个过程等价于解如下的关于 b_i 的方程组：

$$\left\{\begin{aligned} &\sqrt{\sum_{k=1}^{n} x_{1k}^2}\, b_1 + c_{12} b_2 + c_{13} b_3 + \cdots + c_{1s} b_s = r_1 \\ &\qquad \sqrt{\sum_{k=1}^{n} (x_{2k}^{(1)})^2}\, b_2 + c_{23} b_3 + \cdots + c_{2s} b_s = r_2 \\ &\qquad\qquad \sqrt{\sum_{k=1}^{n} (x_{3k}^{(2)})^2}\, b_3 + \cdots + c_{3s} b_s = r_3 \\ &\qquad\qquad\qquad \ddots \\ &\qquad\qquad\qquad\qquad \sqrt{\sum_{k=1}^{n} (x_{sk}^{(s-1)})^2}\, b_s = r_s \end{aligned}\right. \tag{4.4.33}$$

4.4.3 筛选中的一些问题

用正交筛选法选出的 s 个变量，在理论上和实践上都不能保证它们就是最优的。也就是说，可以用别的方法找到另外的 s 个变量(它们与这里选出的 s 个变量至少有一个不相同)，它们构成的 y 的表达式的残差平方和不大于这里选出的 s 个量相应的残差平方和。但大量来自实际的问题的计算结果表明，这种差别往往不是本质的，特别当 s 不太大时，它们往往就是最优的。因此，正交筛选法还是有效的。

正交筛选法的计算程序不难编出，这里从略。但是必须指出，这种计算办法的存贮量和运算量都较大，从节省计算量的角度考虑，正交筛选过程的全部量都可以用递推公式来完成。例如

㊀ 在数理统计中，此量服从 F 分布，自由度为 1 和 $n-s-1$。

$$\sum_{k=1}^{n} x_{ik}^{(s)} y_k = \sum_{k=1}^{n} (x_{ik}^{(s-1)} - c_{si} z_{sk}) y_k = \sum_{k=1}^{n} x_{ik}^{(s-1)} y_k - c_{si} \sum_{k=1}^{n} x_{sk}^{(s-1)} y_k \Big/ \sqrt{\sum_{k=1}^{n} (x_{sk}^{(s-1)})^2} \quad (4.4.34)$$

这样，不难看出式(4.4.33)可以由§4.2的式(4.2.16)用消去法得到。因此，若把选取显著量的思想用于消去过程中(即按 $r_{iy}^{(s)}$ 的大小决定消去顺序，并作显著性检验)，也可以收到同样的效果[7]。本书回归分析一章中给出的逐步回归法正是这样做的。

若式(4.4.30)中的临界值 F_0 很小(特别 $F_0=0$)，则全部变量都被选中。这时 $s=m$，正交筛选法与§4.2指出的传统算法的计算结果在理论上没有差别。但由于算法不同，数字效果也可能不同。此外当自变量之间相关密切，致使法方程出现病态甚至退化时，传统算法将碰到困难，且得到的解 b_i 的精度也不好；而正交筛选法则较自然地回避了这个难点，因为引起麻烦的变量不会被选进，这时有 $s<m$。

最后应该指出，正交筛选法实际用于变量的选取时，被选中的变量个数 s 的大小是值得注意的。只靠统计判断(4.4.29)，(4.4.30)有时并不能选到恰当的 s，还应根据实际问题的特点予以调整。对于 n 较小的问题，s 不应选得太大。特别是为各种预报而建立的表达式，更应注意不要因为片面追求降低残差平方和而选取过大的 s，否则被确定的 b_i 不可靠，往往随着观测数据的微小变化而产生很大的波动。此外，经筛选得到的变量，亦应联系实际问题加以分析。但这种分析要仔细，切忌表面和马虎。例如有些变量，根据经验是非常重要的，但筛选时没有被选中，或放在很不重要的地位，这可能是用于计算的这些变量的观测数据变化甚微，显不出其重要性。这意味着，如果在今后的实际情况中，这些变量只允许作这样微小的变化的话，则 y 的表达式中可以不必考虑它们，或不必作为重要的变量来看待它们。同理，对一些经筛选认为是较次要的量，它们的相应系数 b_i 可能不明显地具有预想的性质(如符号的正、负或数值的大、小等)，这一点也是完全可以预料的。

4.4.4 表达式的半自动挑选

如4.2.2节所述，变量变换的方法可将线性模型大大推广。上述正交筛选法也可以推广到如下模型之中：

$$h(y) = c_1(b_1)g_1(\boldsymbol{X}) + c_2(b_2)g_2(\boldsymbol{X}) + \cdots + c_m(b_m)g_m(\boldsymbol{X}) \quad (4.4.35)$$

其中 $h(y)$ 是因变量的函数，可视作新的因变量；$c_i(b_i)$ 是第 i 个参数 b_i 的函数，可视作新的参数；$g_i(\boldsymbol{X})$ 是变量 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)$ 的函数，可视作新的变量参加正交筛选。由于变量变换的形式是多种多样的，有时还可以是一个变量的各种初等函数的组合或几个变量的相当复杂的组合。因此，对新变量的自动筛选，可代替对于原变量可能列出的形如(4.4.35)的大量表达式的自动挑选。这也是正交筛选法的一种应用。

以一个自变量为例：这时表达式可以从自变量 x 的各种初等函数(如各种幂次、指数、对数、三角函数等)的线性组合中选取。也就是说，可以把

$$\begin{cases} x,\ x^2,\ x^3,\ x^4,\ 1/x,\ 1/x^2,\ \sqrt{x},\ 1/\sqrt{x}, \\ \ln x,\ x\ln x,\ \ln x/x,\ e^x,\ xe^x, \\ \sin x,\ \cos x,\ x\sin x,\ x\cos x, \cdots, \text{等等} \end{cases} \quad (4.4.36)$$

作为各种新的自变量 x_i，用4.4.2节中介绍的正交筛选法选出一个项数不多的表达式来。例如，记为 $y=f_1(x)$。如果有必要，因变量 y 也可以作一些最简单的变换，如

$$\ln y,\ 1/y,\ y^2,\ e^y, \cdots, \text{等等} \quad (4.4.37)$$

对这些新因变量，分别用(4.4.36)所示的新变量作正交筛选建立表达式得

$$\begin{cases} \ln y = f_2(x) & \text{或} \quad y = e^{f_2(x)} \\ 1/y = f_3(x) & \text{或} \quad y = 1/f_3(x) \\ e^y = f_4(x) & \text{或} \quad y = \ln f_4(x) \end{cases} \tag{4.4.38}$$

在这样的基础上，可以根据一定的标准从中选出较好的表达式。

这种途径，完全可以用于本来就有多个自变量的情形。在此情况中，当然还应当考虑这些自变量的某些组合，如

$$x_1x_2,\ x_1/x_2,\ \sqrt{x_1x_2},\ x_1\ln x_2,\ \cdots,\ \text{等等} \tag{4.4.39}$$

上述自动建立表达式的工作，可以编制成通用的计算程序。作为通用程序，式(4.4.36)、(4.4.37)、(4.4.39)的形式都可以更多一些。但应注意变量变换合理：如 $x=0$，则不应有 $1/x$ 的变换；当 $x<0$ 时不能用 $\sqrt{x}$ 等。这些工作，在正交筛选前，程序不难先行处理。例如先把 x 变为 $1\leqslant x\leqslant 2$，其后再作其它变换，一般都不会再遇到困难。

显然，上述的表达式的半自动选取方法，更适用于对表达式形式没有什么限制，而且精度要求不高的一类问题。经验表明，这个方法用于自动控制或一些预报问题所需的表达式，以及数据、曲线、图表公式化等问题时是有效的。

这里附带指出，在建立表达式时，如果实际情况允许，可以使用插值那章介绍的样条函数来建立分段的表达式。这时每段的表达式都可以不太复杂，但总体效果很好。这里不再介绍。

§4.5 随机尝试法

这里将研究最一般的情形。首先表达式

$$y=f(\boldsymbol{X},\ \boldsymbol{B}) \tag{4.5.1}$$

既可以是线性的，也可以是非线性的；其次待定参数可以是在最小二乘意义或其它意义下解出。

先引进一个求极值的模型。设

$$I=F(\boldsymbol{X},\ \boldsymbol{B}) \tag{4.5.2}$$

其中 $\boldsymbol{X}$ 是已知量，它可以是一个向量或矩阵；

$\boldsymbol{B}=(b_1,\ b_2,\ \cdots,\ b_m)$ 是 m 个待定参数；

F 是 $\boldsymbol{X}$、$\boldsymbol{B}$ 的任意非负函数。

问题是求解使 I 为极小的 $\boldsymbol{B}$ 值。在某些时候，还令 $\boldsymbol{B}$ 满足一定的限制条件。

由于 F 的任意性，这个模型实际上包括不少的数学问题，特别包括本章讨论过的曲线拟合中的参数确定问题。例如当表达式为

$$y=f(\boldsymbol{X},\ \boldsymbol{B}) \tag{4.5.3}$$

如果要在最小二乘意义下求解 $\boldsymbol{B}$，则令

$$I=\sum_{k=1}^{n}[y_k-f(\boldsymbol{X}_k,\ \boldsymbol{B})]^2 \tag{4.5.4}$$

此外也可以在非最小二乘意义下确定 $\boldsymbol{B}$，例如求 $\boldsymbol{B}$ 使得

$$I=\max_{1\leqslant k\leqslant n}|y_k-f(\boldsymbol{X}_k,\ \boldsymbol{B})| \tag{4.5.5}$$

或

$$I=\sum_{k=1}^{n}|y_k-f(X_k, B)| \quad (4.5.6)$$

达到最小。

由于 F 的任意性，要给出一般性的有效算法是困难的。尝试法(或称搜索法)是在其它方法不能使用时的一种算法。通常的尝试法是指等距离的网格尝试法，即在需要尝试的区间内插入若干个等分点作为尝试值，并逐一计算这些分点上的 I 值，当全部计算完毕之后，就可选出使 I 值达到最小的那个尝试值作为解。优选法也是一种尝试法，它是对等距离的网格尝试法的一种改进，在待定参数个数较少的情况下是比较有效的。这里介绍一般的随机尝试法及其改进，先从一般的随机尝试法开始。

4.5.1 一般的随机尝试法

对待定参数 b_i 给出一个求解区间(即尝试区间或搜索区间)：

$$\mu_i-\Delta_i\leqslant b_i\leqslant\mu_i+\Delta_i \qquad (i=1, 2, \cdots, m) \quad (4.5.7)$$

每次分别独立地产生 m 个上述区间上的均匀分布随机数 ξ_i㊀，并令

$$\boldsymbol{B}=(b_1, b_2, \cdots, b_m)=(\xi_1, \xi_2, \cdots, \xi_m) \quad (4.5.8)$$

作为一组尝试值，代入下式计算 I：

$$I=F(\boldsymbol{X}, \boldsymbol{B}) \quad (4.5.9)$$

反复产生 $\boldsymbol{B}$，反复尝试(计算 I)，就可以得到如下序列：

$$\boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}, \cdots, \boldsymbol{B}^{(T)}, \cdots \quad (4.5.10)$$

这里的 $\boldsymbol{B}^{(1)}$ 是 $\boldsymbol{B}$ 的第一组尝试值，$\boldsymbol{B}^{(2)}$ 是满足下式中最先出现的一组尝试值 $\boldsymbol{B}$：

$$F(\boldsymbol{X}, \boldsymbol{B})<F(\boldsymbol{X}, \boldsymbol{B}^{(1)}) \quad (4.5.11)$$

或

$$I<I^{(1)}$$

(不满足上式的 $\boldsymbol{B}$ 就被舍弃，不再编号，下同)。$\boldsymbol{B}^{(3)}$ 是满足下式中最先出现的一组 $\boldsymbol{B}$ 值：

$$F(\boldsymbol{X}, \boldsymbol{B})<F(\boldsymbol{X}, \boldsymbol{B}^{(2)}) \quad (4.5.12)$$

其余类推，可知与式(4.5.10)的 $B^{(t)}$ 相应的 $I^{(t)}$ 满足下式：

$$I^{(1)}>I^{(2)}>\cdots>I^{(T)}>\cdots \quad (4.5.13)$$

可以证明，只要 ξ_i 是独立的均匀分布随机数，总能得到使 I 为极小的 $\boldsymbol{B}$ 值。

4.5.2 改进的随机尝试法

这里给出随机尝试法的一种改进——自动收缩求解区间的随机尝试法。显然，尝试区

㊀ 所谓 (a, b) 区间上的独立均匀分布的随机数，通俗地说是指这样的数，它可以在 (a, b) 区间上的任一点取值，而且前一次取值与后一次取值无关(独立性)，但从大量的取值来看，它们并不偏倚于 (a, b) 区间上的某一局部位置，而是均匀地分布的。这样的随机数可以在电子计算机上用递推的方法近似地得到。有些计算机把产生随机数作为标准子程序与其它初等函数的标准子程序一起给出。

下面介绍一种产生方法：例如要产生 $(0, 2^M)$ 区间上的独立均匀随机数，可用"乘同余法"递推得到：

$$\eta_{n+1}=\eta_n\cdot\lambda(\mathrm{mod}\,2^M)$$

这里 M 为机器的尾数位数。$\mathrm{mod}2^M$ 即取 $\eta_n\cdot\lambda$ 的乘积的后 M 位。η_0 可取任一奇数，例如 $\eta_0=1, 3, 5, \cdots$，等。λ 是不超过 2^M 的常数，一般可取 $\lambda=5^{2k+1}$(k 是正整数，它使 λ 尽可能靠近 2^M)。在机器上，如果把最后一位二进制数看作是 1，则 η_n 与 λ 的乘法可用"不规舍"的双倍位乘法指令实现，取模即取此乘积的后 M 位。在使用时，仍依习惯，即把最后一位二进制数看作为 2^{-M}，则 $(0, 2^M)$ 区间上的随机数即为 $(0, 1)$ 区间上的随机数。有了 $(0, 1)$ 区间的随机数 η，就不难得到 (a, b) 区间上的均匀随机数 ξ：$\xi=a+\eta(b-a)$。

间的大小,直接影响求解效率,如果在尝试过程中,充分利用前面计算所带来的信息,不断缩小求解区间,将有可能大大节省计算时间。下面介绍的正是这样的一种方法,具体步骤是:

(1) 指定一个正整数 T(例如 $T=5$),仿照 4.5.1 节求得 T 组 $\boldsymbol{B}$ 值:

$$\boldsymbol{B}^{(1)}\ \ \boldsymbol{B}^{(2)},\ \cdots,\ \boldsymbol{B}^{(T)} \tag{4.5.14}$$

如 4.5.1 节所述,与这些 $\boldsymbol{B}^{(t)}$ 相应的 $I^{(t)}$ 值满足

$$I^{(1)}>I^{(2)}>\cdots>I^{(T)} \tag{4.5.15}$$

(2) 指定一个正数 c(例如 $c=2.5$)及一组权 $W^{(t)}$(例如 $W^{(t)}=I^{(T)}/I^{(t)}(t=1,\ 2,\cdots,\ T)$),这组权应满足下述关系:

$$W^{(1)}<W^{(2)}<\cdots<W^{(T)} \tag{4.5.16}$$

在此基础上计算新的求解区间的中点 μ_i 及半长 Δ_i:

$$\mu_i=\sum_{t=1}^{T}W^{(t)}b_i^{(t)}\Big/\sum_{t=1}^{T}W^{(t)}\quad(i=1,\ 2,\ \cdots,\ m) \tag{4.5.17}$$

$$\begin{aligned}\Delta_i&=c\sqrt{\sum_{t=1}^{T}W^{(t)}(b_i^{(t)}-\mu_i)^2\Big/\sum_{t=1}^{T}W^{(t)}}\\&=c\sqrt{\left(\sum_{t=1}^{T}W^{(t)}(b_i^{(t)})^2\Big/\sum_{t=1}^{T}W^{(t)}\right)-\mu_i^2}\qquad(i=1,\ 2,\ \cdots,\ m)\end{aligned} \tag{4.5.18}$$

一般而言,新的区间将有可能具有我们所期待的性质:即 μ_i 比原来的更接近 b_i 的真解,Δ_i 比原来的大为缩小。

(3) 以新的 μ_i、Δ_i 代替原来的 μ_i、Δ_i,并重复上述计算步骤(1)、(2),直至下述条件得到满足:

$$\max_{1<i<m}\left(\Delta_i/(|\mu_i|+0.001)\right)<\varepsilon_1 \tag{4.5.19}$$

或

$$I^{(t)}<\varepsilon_2 \tag{4.5.20}$$

其中 ε_1、ε_2 均为给出的允许误差。

4.5.3 在实际计算中应注意的事项

自动收缩求解区间的随机尝试法是一种经验性和技巧性的方法,因此下面有必要对实际计算中应注意的事项稍加说明:

(1) T 的大小值得注意。过大起不到收缩区间的作用,过小又容易把求解"引入死胡同"(即正在尝试的求解区间已不包含真解了)。第一轮收缩时,T 例如可取 4、5、6 等值。其后各轮收缩时,T 值不变。但为了提高效率只要有了两组新 $\boldsymbol{B}$ 值,就可收缩,不足的 $T-2$ 组 $\boldsymbol{B}$ 值,可用上一轮最后出现的补齐。

(2) $\boldsymbol{B}^{(1)}$ 不真正取 B 的第一组尝试值,改为取前 K 次尝试中的最优值(例如 $K=30\mathrm{m}$)。

(3) 上述算法有可能导致新的尝试区间越出最早给出的区间(当 c 较大时,更易出现),最终解也因而可能在最早给出的区间之外求得。这是本算法的一个优点。但是对于解的范围有一定限制的问题,应稍加处理。

(4) 如果在某组 $\boldsymbol{B}^{(t)}$ 之后,虽经很多次尝试,仍不能获得 $\boldsymbol{B}^{(t+1)}$,则可改令 $\mu=\boldsymbol{B}^{(t)}$,当前的 Δ 的 1/5 作为新的 Δ,强行收缩。

(5) 尝试区间要合理，以免造成计算 I 时的困难(如溢出)。此外，在曲线拟合问题中，I 值将随着观测顺序号 k 的增加而增加(不减)，因此，在每次尝试中，不必等待 n 个观测点算完便可判别这次的 I 值是否大于当前最小的 I 值。也就是说，算一个点，判别一次。这样可以大大节省计算时间。经验表明，当尝试次数较多时，不论 n 值有多大，每次尝试所需计算的观测点数，平均而言不超过 2。

(6) 对于给出的尝试区间 $(\mu-\varDelta, \mu+\varDelta)$，一般总认为解落在该区间的中间部分的可能性大。独立均匀分布的随机数不能反映这一点。为此，可以交替使用独立的均匀分布随机数和独立的三角分布随机数㊀。

这里提供的方法，既考虑到提高计算效率，也考虑到不要把解“引入死胡同”。当然，不论在理论上还是实践上都不能保证这种算法一定成功。特别是对于有多解而且各种解靠得很近，或 $F(\boldsymbol{X}, \boldsymbol{B})$ 对于 $\boldsymbol{B}$ 而言变化激烈的问题，随机尝试法是难以收到良好效果的。

随机尝试法的算法简单，通用性强，对于曲线拟合中各种意义下确定 $\boldsymbol{B}$ 都同样方便，而且对 $\boldsymbol{B}$ 附加若干限制也易于处理。但一般而言，它耗费机器时间较多，仅作为应急或其它数值方法无效时之替代手段。计算经验表明，对于重复计算次数较小的问题，如果对其它算法的把握不大，或没有其它现成的算法程序，则采用随机尝试法，有可能在一次上机计算中花费不多的时间取得结果，从而大大节省人的劳动和缩短解题周期。对于精度要求较高的问题，本法常常可以提供一个较好的初值。

下面是两个数字例子：

例 1　曲线拟合问题：

给出 10 个观测点 (x_k, y_k) $(k=1, 2, \cdots, 10)$，并指定表达式的形式为 $y=b_1+b_2x$，要求 b_1、b_2 使 $I=\max\limits_{1<k<10}|y_k-(b_1+b_2x_k)|$ 达到极小。

此题的真解已知为 $b_1=1$, $b_2=2$，在利用随机尝试法(有收缩区间)求解时，故意令真解落在给出的求解区间的边缘上，其中 b_1 的求解区间为(1, 3)，即 $\mu_1=2$, $\varDelta_1=1$；b_2 的区间为(0, 2)，即 $\mu_2=1$, $\varDelta_2=1$。在 109-乙机上经过不足 2000 次尝试(3 秒钟)获得的结果与真解的前 5 位相同。

例 2　解非线性代数方程组：

下列三个未知数的非线性代数方程组来源于高温燃气热力计算：

$$\begin{cases}f_1(x_1, x_2, x_3)=a_7(x_1+x_3)+2x_2+2P_5+P_6+P_7+P_9=0\\ f_2(x_1, x_2, x_3)=(1-a_8)x_1+x_2+(2-a_8)x_3+2P_4+P_6+P_9+P_{10}=0\\ f_3(x_1, x_2, x_3)=x_1+x_2+x_3+P_4+P_5+P_6+P_7+P_8+P_9+P_{10}-a_{10}=0\end{cases}$$

其中 a_i　$(i=1, 2, \cdots, 10)$ 为已知常数；

P_i　$(i=4, 5, \cdots, 10)$ 都是 x_1、x_2、x_3 的函数，具体形式是：

$$P_4=a_1^2(x_3/x_1)^2 \qquad P_5=x_1x_2/a_2x_3$$

㊀ 例如，$(-1, 1)$ 区间上独立的三角分布随机数 η，是指这样的随机数：它可以在 $(-1, 1)$ 区间上任意一点取值，而且前一次取值与后一次取值无关(独立性)，但从大量的取值看，它偏倚该区间的中部，即取值最多的是 0 附近，离 0 越远，取值可能越小，其频率呈一个等腰三角形状态。

这样的随机数可用二个独立均匀的随机数 η_1 和 η_2 [都是 $(0, 1)$ 区间上的] 相减得．$\eta=\eta_1-\eta_2$。从 $(-1, 1)$ 上的三角分布化为 (a, b) 上的三角分布 ξ，可用变换 $\xi=a+\frac{(b-a)}{2}(\eta+1)$。

$$P_6=a_3\sqrt{a_2}\sqrt{\frac{x_2x_3}{x_1}}\qquad P_7=\sqrt{\frac{a_4}{a_2}}\sqrt{\frac{x_1x_2}{x_3}}$$

$$P_8=a_6a_9(x_1+x_3)\Big/\left(a_6+a_3\sqrt{a_2}\sqrt{\frac{x_2x_3}{x_1}}\right)$$

$$P_9=a_9(x_1+x_3)\Big/\left(1+1/a_3\sqrt{a_2}\sqrt{\frac{x_2x_3}{x_1}}\right)$$

$$P_{10}=a_1\sqrt{a_5}\,x_3/x_1$$

在利用随机尝试法(有收缩区间)求解时，令 $I=f_1^2+f_2^2+f_3^2$，尝试区间都是(0, 0.6)，即 $\mu_i=0.3$, $\Delta_i=0.3$ $(i=1, 2, 3)$。在 109-乙机上经 14300 次尝试(30 秒)得解:

$$x_1=0.1416475\qquad x_2=0.5114150\qquad x_3=0.3770055$$

与真解的前 4 位相同。类似的方程解过 30 个,效果近似(尝试次数都在 2 万次左右)。

参 考 资 料

[1] Marquardt, D. W., "An algorithm for least-squares estimation of nonlinear parameters", JSIAM Vol. 11, No. 2, p. 431–441, 1963.

[2] Powell, M. JD., "A method for minimizing a sun of squares of nonlinar functions without calculating derivatives", The Computer Journal Vol. 7, p. 303–307, 1965.

[3] Brown, K. M. Dennis, T. E., "Derivative free anoloques of the Levenberg-Marquardt and Gauss algorithm for nonlinear least squares approximation", Numerische Mathematik 18 Band 4 Heft, p. 289–297, 1972.

[4] Jones, A., "Spinal—A new algorithm for nonlinear parameters estimation using least-squares", The Computer Journal Vol. 13, No. 3, p. 301–308, 1970.

[5] Shanno, D. F., "Parameter selection for modified Newton methods for function minimization", SIAM Journal on Numerical analysis Vol. 7, No. 3, p. 366–372, 1970.

[6] Aubert, E. J. et al, "Some objective six-hour predictions prepared by statistical methods": Journal of Meteorology. Vol. 16, No. 4, p. 436–446, 1959.

[7] A. 拉斯登等著，徐献瑜等译，《数字计算机上用的数学方法》，第 17 章，上海科技出版社，1963。

第五章 回归分析

§5.1 回归问题

回归分析计算的中心问题，就是根据由变量组

$$(x_1,\ x_2,\ \cdots,\ x_m;\ y)$$

得到的 N 组观测数据

$$(x_{n1},\ x_{n2},\ \cdots,\ x_{nm};\ y_n) \tag{5.1.1}$$

$$n=1,\ 2,\ \cdots,\ N,\quad N>m$$

对线性回归方程

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m+\varepsilon=\beta_0+\sum_{i=1}^{m}\beta_ix_i+\varepsilon \tag{5.1.2}$$

进行最佳拟合，研究因变量 y 和自变量 $x_1,\ x_2,\ \cdots,\ x_m$ 之间的关系，预报或控制因变量 y 的取值。

在实际问题中，假定 N 组观测数据(5.1.1)满足线性回归方程(5.1.2)，即

$$y_n=\beta_0+\sum_{i=1}^{m}\beta_ix_{ni}+\varepsilon_n \tag{5.1.3}$$

$$n=1,\ 2,\ \cdots,\ N$$

并把因变量 y 叫做预报量，$x_1,\ x_2,\ \cdots,\ x_m$ 叫做预报因子。

在回归方程(5.1.2)、(5.1.3)中，回归系数 $\beta_0,\ \beta_1,\ \cdots,\ \beta_m$，是未知的待定参量，$\varepsilon$ 是随机变量，表示因变量 y 去掉一般变量 $x_1,\ x_2,\ \cdots,\ x_m$ 的影响后，由许多复杂或然因素存在形成的随机误差。一般假定各次观测过程中出现的随机误差 $\varepsilon_n(n=1,\ 2,\ \cdots,\ N)$ 相互独立，服从数学期望为零、方差为 σ^2 的正态分布。为简单计，今后把数学期望为 μ、方差为 σ^2 的正态分布，简记为 $N(\mu,\ \sigma^2)$。例如，将数学期望为零，方差为 σ^2 的正态随机误差 ε，简记为 $\varepsilon\sim N(0,\ \sigma^2)$。

现在，我们用一个简单例子，说明上面提到的一些概念。

在用转炉炼钢过程中，每炉钢的钢水重量要求给于预报和控制，取为回归方程中的预报量 y。炼钢过程中，影响钢水重量的一些可控因素，如兑入转炉里的铁水重量，废钢、矿石、石灰石等副原料的加入量，供氧的情况等，取作自变量。要求我们研究它们之间的关系，建立进行预报和控制的回归模型(5.1.2)。炼钢过程中的观测误差和其它一些或然误差，作为回归方程中的误差项 ε。在每个炉役中，除一些可控因素外，每炉钢都是在大致类似的条件下冶炼的。炼钢过程中得到的观测数据，有着大致相同的精度。根据上述诸点，可合理地假定误差项 ε_n 相互独立地服从 $N(0,\ \sigma^2)$ 分布。

在回归分析计算中，要解决的主要问题有：

(1) 根据给出的 N 组观测数据(5.1.1)，给出回归系数 $\beta_0,\ \beta_1,\ \cdots,\ \beta_m$ 的最小二乘估计值 $b_0,\ b_1,\ \cdots,b_m$(参见§5.2)，定量表示 y 和 $x_1,\ x_2,\ \cdots,\ x_m$ 之间的关系。对估计值 b_i 进

行统计检验,给出这一估计的可靠性。

(2)根据自变量$(x_1, x_2, \cdots, x_m)$的一组观测值,预报因变量y的取值,并且给出预报的精度,即给出ε未知方差σ^2的估计值。

回归分析是一类应用范围很广的统计分析方法。它在一般数据处理、产品质量控制、建立自动控制数学模型、曲线拟合,以及气象、水文、地震等预报中,都有着重要的应用[3~5]。

§5.2 法 方 程

给出线性回归方程(5.1.2)中回归系数$\beta_0, \beta_1, \cdots, \beta_m$的估计值$b_0, b_1, \cdots, b_m$,记

$$y_n^*=b_0+b_1x_{n1}+b_2x_{n2}+\cdots+b_mx_{nm}=b_0+\sum_{i=1}^m b_ix_{ni} \tag{5.2.1}$$

叫做y_n的预报值或估计值,它们之差

$$e_n=y_n-y_n^*=y_n-\left(b_0+\sum_{i=1}^m b_ix_{ni}\right) \tag{5.2.2}$$

叫做预报残差。

最小二乘估计的原则是,定出估计值$b_0, b_1, \cdots, b_m$,使残差平方和

$$Q=\sum_{n=1}^N e_n^2=\sum_n [y_n-(b_0+\sum_i b_ix_{ni})]^2 \tag{5.2.3}$$

最小。今后,在不致发生混淆时,记

$$\sum_{n=1}^N e_n^2=\sum_n e_n^2$$

$$\sum_{i=1}^m b_ix_{ni}=\sum_i b_ix_{ni}$$

对给定的N组观测数据$(x_{n1}, x_{n2}, \cdots, x_{nm}; y_n)$,残差平方和$Q$是待定参量$b_0, b_1, \cdots, b_m$的二次函数,非负,最小值存在。根据数学分析中的极值原理,选取$b_0, b_1, \cdots, b_m$使Q最小的条件是:

$$\frac{\partial Q}{\partial b_i}=0$$

对$i=0, 1, \cdots, m$成立,即Q对b_i的偏导数全部为0。

由$\dfrac{\partial Q}{\partial b_0}=0$,得到

$$\sum_n [y_n-b_0-\sum_i b_ix_{ni}]=0$$

即

$$\sum_n e_n=0$$

$$b_0=\frac{1}{N}\sum_n y_n-\sum_i b_i\left(\frac{1}{N}\sum_n x_{ni}\right)=\bar{y}-\sum_i b_i\bar{x}_i \tag{5.2.4}$$

这里

$$\bar{y}=\frac{1}{N}\sum_n y_n$$

$$\bar{x}_i=\frac{1}{N}\sum_n x_{ni} \quad (i=1, 2, \cdots, m)$$

分别称为预报量y和预报因子x_i的均值。把(5.2.4)代入(5.2.3),整理后,得到残差平方和

$$Q=\sum_{n}[(y_n-\bar{y})-\sum_{i}b_i(x_{ni}-\bar{x}_i)]^2 \tag{5.2.5}$$

根据 $\frac{\partial Q}{\partial b_i}=0$, 在 $i=1, 2, \cdots, m$ 时,有

$$\sum_{n}[(y_n-\bar{y})-\sum_{j}b_j(x_{nj}-\bar{x}_j)](x_{ni}-\bar{x}_i)=0$$

即

$$\sum_{n}e_n(x_{ni}-\bar{x}_i)=0 \tag{5.2.6}$$

由此可得

$$\sum_{j}b_j\sum_{n}(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)=\sum_{n}(x_{ni}-\bar{x}_i)(y_n-\bar{y})$$

记叉乘和

$$l_{ij}=\sum_{n}(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)$$

$$l_{iy}=\sum_{n}(x_{ni}-\bar{x}_i)(y_n-\bar{y})$$

得法方程

$$\begin{cases} l_{11}b_1+l_{12}b_2+\cdots+l_{1m}b_m=l_{1y} \\ l_{21}b_1+l_{22}b_2+\cdots+l_{2m}b_m=l_{2y} \\ \cdots\cdots \\ l_{m1}b_1+l_{m2}b_2+\cdots+l_{mm}b_m=l_{my} \end{cases} \tag{5.2.7}$$

因为 $l_{ij}=l_{ji}$, 法方程(5.2.7)的系数矩阵

$$(l_{ij})=\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \cdots\cdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{pmatrix} \tag{5.2.8}$$

是对称的。如果(l_{ij})可逆, (5.2.7)有唯一解

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}=(l_{ij})^{-1}\begin{pmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{my} \end{pmatrix}=(l^{ij})\begin{pmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{my} \end{pmatrix} \tag{5.2.9}$$

这里记

$$(l^{ij})=(l_{ij})^{-1}$$

根据(5.2.5)、(5.2.6),残差平方和

$$\begin{aligned} Q&=\sum_{n}[(y_n-\bar{y})-\sum_{i}b_i(x_{ni}-\bar{x}_i)]^2 \\ &=\sum_{n}(y_n-\bar{y})[(y_n-\bar{y})-\sum_{i}b_i(x_{ni}-\bar{x}_i)] \\ &\quad -\sum_{i}b_i\sum_{n}[(y_n-\bar{y})-\sum_{j}b_j(x_{nj}-\bar{x}_j)](x_{ni}-\bar{x}_i) \\ &=\sum_{n}(y_n-\bar{y})^2-\sum_{i}b_i\sum_{n}(x_{ni}-\bar{x}_i)(y_n-\bar{y}) \\ &\quad -\sum_{i}b_i\sum_{n}e_n(x_{ni}-\bar{x}_i) \\ &=l_y^2-\sum_{i}b_il_{iy} \end{aligned} \tag{5.2.10}$$

其中

$$l_y^2=\sum_n (y_n-\bar{y})^2$$

当 $N\gg m$ 时，利用(5.2.10)，可以更快地算出 Q 的数值。

§5.3 法方程解的统计性质

这一节讨论§5.2中给出的各个估计量 y_n^*, b_i, Q 的统计性质，给出它们的数学期望、方差和协方差，从而给出它们的估计精度和进行统计检验的方法。这里得到的结果，构成了回归分析的理论基础。它不仅加深我们对回归分析的理解，而且给出了后面将要介绍的逐步回归算法的理论依据[1,2,9]。对只想尽快掌握逐步回归算法或只用回归模型解决诸如曲线拟合一类问题的读者，可越过这一节；需要时，再返回来研究。

根据§5.1中的假定，线性回归模型

$$y_n=\beta_0+\sum_i \beta_i x_{ni}+\varepsilon_n$$

中的自变量 x_{ni} 是一般变量，ε_n 是随机变量，相互独立，服从 $N(0, \sigma^2)$ 分布。所以，因变量 y_n 是正态随机变量，它的数学期望、方差和协方差分别为：

$$\left\{\begin{aligned} &E(y_n)=\beta_0+\sum_i \beta_i x_{ni} \\ &D(y_n)=E[y_n-E(y_n)]^2=\sigma^2 \\ &\operatorname{cov}(y_{n_1}, y_{n_2})=E[(y_{n_1}-E(y_{n_1}))(y_{n_2}-E(y_{n_2}))]=E(\varepsilon_{n_1}\cdot\varepsilon_{n_2})=0 \quad (n_1\neq n_2) \end{aligned}\right. \tag{5.3.1}$$

在法方程(5.2.7)中，系数

$$l_{ij}=\sum_n (x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)$$

是常量；自由项

$$l_{iy}=\sum_n (x_{ni}-\bar{x}_i)(y_n-\bar{y})$$

是正态随机变量 y_n 的线性组合，所以，也是正态的。因此，由(5.2.9)得到的正规方程的解

$$b_i=\sum_j l^{ij} l_{jy} \quad (i=1, 2, \cdots, m)$$

是正态随机变量，且有

$$\left.\begin{aligned} &E(b_i)=\beta_i \\ &\operatorname{cov}(b_i, b_j)=l^{ij}\sigma^2 \end{aligned}\right\} \tag{5.3.2}$$

[证明]：因为

$$\bar{y}=\frac{1}{N}\sum_n y_n=\beta_0+\sum_i \beta_i \bar{x}_i+\bar{\varepsilon}$$

其中
$$\bar{\varepsilon}=\frac{1}{N}\sum_n \varepsilon_n$$

所以
$$\begin{aligned} l_{jy}&=\sum_n (x_{nj}-\bar{x}_j)(y_n-\bar{y})=\sum_n (x_{nj}-\bar{x}_j)\left[\sum_k \beta_k(x_{nk}-\bar{x}_k)+(\varepsilon_n-\bar{\varepsilon})\right] \\ &=\sum_k l_{jk}\beta_k+\sum_n (x_{nj}-\bar{x}_j)(\varepsilon_n-\bar{\varepsilon}) \end{aligned}$$

$$\begin{aligned} b_i&=\sum_j l^{ij} l_{jy}=\sum_k\sum_j l^{ij} l_{jk}\beta_k+\sum_j l^{ij}\sum_n (x_{nj}-\bar{x}_j)(\varepsilon_n-\bar{\varepsilon}) \\ &=\sum_k \delta_{ik}\beta_k+\sum_j l^{ij}\sum_n (x_{nj}-\bar{x}_j)(\varepsilon_n-\bar{\varepsilon}) \\ &=\beta_i+\sum_j l^{ij}\sum_n (x_{nj}-\bar{x}_j)(\varepsilon_n-\bar{\varepsilon}) \end{aligned}$$

这里
$$\delta_{ik}=\begin{cases}1, & \text{当 } i=k \text{ 时}\\ 0, & \text{当 } i\neq k \text{ 时}\end{cases}$$

由于
$$E(\varepsilon_n-\bar{\varepsilon})=0$$

$$E[(\varepsilon_{n_1}-\bar{\varepsilon})(\varepsilon_{n_2}-\bar{\varepsilon})]=\left(\delta_{n_1n_2}-\frac{1}{N}\right)\sigma^2$$

故得
$$E(b_i)=\beta_i$$

$$\begin{aligned}\operatorname{cov}(b_i,\ b_j)&=E[(b_i-\beta_i)(b_j-\beta_j)]\\&=E[\sum_{k_1}l^{ik_1}\sum_{n_1}(x_{n_1k_1}-\bar{x}_{k_1})(\varepsilon_{n_1}-\bar{\varepsilon})\cdot\sum_{k_2}l^{jk_2}\sum_{n_2}(x_{n_2k_2}-\bar{x}_{k_2})(\varepsilon_{n_2}-\bar{\varepsilon})]\\&=\sum_{k_1,k_2}l^{ik_1}l^{jk_2}\sum_{n_1,n_2}(x_{n_1k_1}-\bar{x}_{k_1})(x_{n_2k_2}-\bar{x}_{k_2})E[(\varepsilon_{n_1}-\bar{\varepsilon})(\varepsilon_{n_2}-\bar{\varepsilon})]\\&=\sum_{k_1,k_2}l^{ik_1}l^{jk_2}\sum_{n_1,n_2}(x_{n_1k_1}-\bar{x}_{k_1})(x_{n_2k_2}-\bar{x}_{k_2})\left(\delta_{n_1n_2}-\frac{1}{N}\right)\sigma^2\\&=\sum_{k_1,k_2}l^{ik_1}l^{jk_2}l_{k_1k_2}\sigma^2=l^{ij}\sigma^2\qquad\text{证毕}\end{aligned}$$

根据(5.2.4)
$$b_0=\bar{y}-\sum_i b_i\bar{x}_i=\beta_0+\sum_i(\beta_i-b_i)\bar{x}_i+\bar{\varepsilon}$$

有残差
$$\begin{aligned}e_n=y_n-y_n^*&=\beta_0+\sum_i\beta_i x_{ni}+\varepsilon_n-b_0-\sum_i b_i x_{ni}\\&=(\varepsilon_n-\bar{\varepsilon})-\sum_i(b_i-\beta_i)(x_{ni}-\bar{x}_i)\end{aligned}$$

故得
$$E(e_n)=0$$

$$\begin{aligned}E(e_n^2)&=E(\varepsilon_n-\bar{\varepsilon})^2+\sum_{i,j}(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)E[(b_i-\beta_i)(b_j-\beta_j)]\\&\quad-2\sum_i(x_{ni}-\bar{x}_i)E[(\varepsilon_n-\bar{\varepsilon})(b_i-\beta_i)]\\&=\left(1-\frac{1}{N}\right)\sigma^2+\sum_{i,j}(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)l^{ij}\sigma^2\\&\quad-2\sum_i(x_{ni}-\bar{x}_i)\sum_j l^{ij}\sum_{n_1}(x_{n_1j}-\bar{x}_j)\left(\delta_{nn_1}-\frac{1}{N}\right)\sigma^2\\&=\left[\frac{N-1}{N}-\sum_{i,j}(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)l^{ij}\right]\sigma^2\end{aligned}$$

$$\begin{aligned}E(Q)=\sum_n E(e_n^2)&=[N-1-\sum_{i,j}l^{ij}\sum_n(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)]\sigma^2\\&=(N-1-\sum_{i,j}l^{ij}l_{ji})\sigma^2=(N-m-1)\sigma^2\end{aligned}$$

最后得到 y、ε 方差 σ^2 的无偏估计
$$\bar{\sigma}^2=\frac{Q}{N-m-1}\tag{5.3.3}$$

根据(5.3.1~5.3.3)，我们得到下面三个重要的结果：

(1) 预报量 y 的估计精度

对自变量 $(x_1, x_2, \cdots, x_m)$ 的一组给定值，根据(5.3.1)，y 服从 $N(\beta_0+\sum_i\beta_i x_i,\ \sigma^2)$ 分布。由(5.3.2)、(5.3.3)可知，预报值 y 围绕着
$$y^*=b_0+\sum_i b_i x_i$$

对称分布取值，愈靠近 y^*，取值的可能性愈大。

记 y 的取值范围为

$$(y^*-\lambda_\alpha\bar{\sigma},\ y^*+\lambda_\alpha\bar{\sigma})$$

它和 λ_α 有如下的渐近统计关系：

λ_α	y 的取值概率 α
0.5	0.383
0.6745	0.500
1.0	0.683
2.0	0.955
3.0	0.997

因 $\bar{\sigma}=\sqrt{\dfrac{Q}{N-m-1}}$，所以 y 的估计精度与残差平方和 Q 以及回归方程中预报因子的个数 m 的平方根成反比。在用回归模型进行预报或控制时，应选用尽可能少的预报因子，达到尽可能高的拟合度，即要求 m、Q 都尽可能小。

我们知道，在回归方程中，增加一个预报因子，残差平方和 Q 将减少，而减少一个预报因子，Q 将增加。因此，同时要求 Q 和 m 都很小，这是矛盾的。为了在这两个相互矛盾的要求下，得到最优的回归方程，对给出的预报因子应择优使用，有所舍选。为此，给出预报因子舍选的统计检验方法。

(2) 回归系数 β_i 的统计检验

由(5.3.2)可知，法方程的解 b_i 是回归系数 β_i 的无偏估计，服从 $N(\beta_i,\ l^{ii}\sigma^2)$ 分布。因为 $\mathrm{cov}(b_i,\ e_n)=0$，故根据(5.3.3)，统计量

$$\frac{(N-m-1)(b_i-\beta_i)^2}{l^{ii}\cdot Q}$$

服从自由度 $(1,\ N-m-1)$ 的 F-分布。特别，当

$$F_i=\frac{(N-m-1)b_i^2}{l^{ii}Q}<F_\alpha \tag{5.3.4}$$

时，回归系数 β_i 接近于 0，这时，可把预报因子 x_i 从回归方程(5.1.2)中舍去。这里，F_α 是在显著水平 α 给定时，由 F-分布表中查得的一个常数。可以证明

$$U_i^2=b_i^2/l^{ii} \tag{5.3.5}$$

是预报因子 x_i 在降低残差平方和 Q 中的贡献，即由回归方程中剔除预报因子 x_i 后，残差平方和 Q 的增加量。今后，简称 U_i^2 为预报因子 x_i 的贡献。统计检验(5.3.4)告诉我们，把贡献不显著的预报因子从回归模型中剔除出去，是合理的。

(3) 回归问题的残差分析和回归方程的 F-检验

对于得到的回归方程

$$y^*=b_0+\sum_i b_ix_i$$

象分析 b_i 一样，也需要给出判别好坏，能否接受的统计检验方法。为此，需要对残差平方和作进一步的分析。

根据(5.2.10)，残差平方和

$$Q=l_y^2-\sum_i b_i l_{iy}=l_y^2-U^2 \tag{5.3.6}$$

其中

$$l_y^2=\sum_n (y_n-\bar{y})^2$$

叫做总平方和，由给出的 N 组观测数据(5.1.1)完全确定。

$$U^2=\sum_i b_i l_{iy}=\sum_i b_i\sum_j b_j l_{ij}=\sum_{i,j} b_i b_j \sum_n (x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j)$$

$$=\sum_n \left[\sum_i b_i (x_{ni}-\bar{x}_i)\right]^2=\sum_n (y_n^*-\bar{y})^2$$

称为回归平方和，是引进预报因子$(x_1, x_2, \cdots, x_m)$后，总平方和 l_y^2 的降低量，反映了预报因子的共同作用。

很显然，在回归分析中，要使 Q 尽可能的小，就必须要求 U^2 尽可能的接近 l_y^2。因此，Q、l_y^2、U^2 之间的相对大小，可用来说明回归方程的好坏。

取无量纲统计量

$$R=\frac{U}{l_y}=\left[\frac{\sum_n (y_n^*-\bar{y})^2}{\sum_n (y_n-\bar{y})^2}\right]^{1/2} \tag{5.3.7}$$

作为回归方程好坏的第一个标准。因为

$$R=\frac{\sum_n (y_n-\bar{y})(y_n^*-\bar{y})}{\sqrt{\sum_n (y_n-\bar{y})^2\cdot\sum_n (y_n^*-\bar{y})^2}}$$

表示自变量$(x_1, x_2, \cdots, x_m)$和因变量 y 的相关，所以称 R 为复相关系数。显然，$0\leqslant R\leqslant 1$，愈接近于1，回归方程的拟合度愈好。

取

$$F=\frac{(N-m-1)(l_y^2-Q)}{mQ} \tag{5.3.8}$$

为回归方程好坏的第二个标准。根据(5.3.3)、(5.3.6)，统计量 F 服从自由度$(m, N-m-1)$的 F-分布。和统计检验(5.3.4)相比，(5.3.8)给出了一组预报因子能否接受的统计检验方法。

§5.4 预报因子舍选和逐步回归计算

在回归分析的实际应用中，总是先选取和预报量 y 多少有一定关系的一组变量作为可能的预报因子。比如用回归模型进行气象预报时，就选有各种不同的气象要素，如温度、湿度、气压、风向、风速等等和不同台站对这些气象要素的观测数据以及它们之间的若干组合，作为可能预报因子。这种可能预报因子的数目常达数十个到近百个。理论分析和实际经验告诉我们，把给出的全部预报因子不经过统计检验的舍选，全部放入回归方程，往往导致法方程(5.2.7)的系数矩阵(l_{ij})蜕化，无法求解，或解得的回归方程精度不高，实际中无法应用。因此，对给出的可能预报因子，必须根据它们在回归方程中的贡献大小，选入回归方程。为此，需要进一步分析各个预报因子在回归方程中的作用。

令 $Q^{(m)}$ 表示由 m 个预报因子$(x_1, x_2, \cdots, x_m)$组成回归方程时的残差平方和，$Q_i^{(m-1)}$ 表示剔除一个预报因子 $x_i(1\leqslant i\leqslant m)$后，由$(m-1)$个预报因子组成回归方程时的残差平方和，

则

$$U_i^2=Q_i^{(m-1)}-Q^{(m)}$$

是预报因子 x_i 在 m 个因子组成回归方程时降低残差平方和的贡献。显然，$U_i^2\geqslant 0$，且

$$U_i^2=b_i^2/l^{ii}$$

因 b_i 和 l^{ii} 随回归方程中预报因子的变化和个数的不同而异，所以，每个预报因子对回归方程降低残差的贡献是不同的，而且是变化的。这就要求我们，不单要从那些尚未选入回归方程的待选预报因子中，选取具有最大贡献的因子进入回归方程[4]，而且还要考虑那些已经进入回归方程的已选预报因子的贡献是否显著。假若选入回归方程的预报因子，在回归方程又选入一些别的因子之后，贡献发生变化，不再显著，则在选入新的预报因子之前，应把贡献最差的已选因子，从回归方程中剔除出去。这样，才能使我们最后得到的回归方程，只包含一些贡献显著的预报因子。我们把满足上述要求的回归算法，称为逐步回归。

下面，根据统计检验(5.3.4)对法方程(5.2.7)，给出适合上述要求的逐步回归算法[1,9]。

为了更清楚地说明逐步回归算法，记法方程(5.2.7)为

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots\cdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \cdots\cdots & & & \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{pmatrix}\begin{pmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{my} \end{pmatrix}$$

或用矩阵符号简记为

$$(a_{ij})(b_j)=(c_{ij})(l_{jy}) \tag{5.4.1}$$

开始计算时，取$(a_{ij})=(l_{ij})$，这时$(c_{ij})=(\delta_{ij})$，是一个 m 阶的单位矩阵。用逐步回归算法求解(5.4.1)的过程，就是通过 b_j 的一步一步的消元变换，变(a_{ij})为(δ_{ij})、(c_{ij})为$(l^{ij})=(l_{ij})^{-1}$的过程。

在(5.4.1)中，如果把 l_{jy} 也看作未知量，方程两边的系数矩阵(a_{ij})和(c_{ij})、未知向量(b_j)和(l_{jy})具有明显的对称性，但它们的意义和作用是完全不同的。在(a_{ij})中，对 b_j 进行消元变换，相当于把待选预报因子选入回归方程，是法方程(5.2.7)的消元求解；在(c_{ij})中，对 l_{jy} 进行消元变换，相当于把已选预报因子从回归方程中剔除出去，是回归方程的消元求解。这样，利用对 l_{jy} 和 b_j 的两类消元变换，恰好可以实现预报因子的舍选。由于它们之间的对称性，这两类消元过程在计算上是完全一致的，有利于在数字计算机上进行处理。在不增加多少运算量的基础上，利用这种算法，还可以得到一系列的过渡性的回归方程，如：

$$y^{(1)}=b_0^{(1)}+b_1^{(1)}x_{i_1}$$

$$y^{(2)}=b_0^{(2)}+b_1^{(2)}x_{i_1}+b_2^{(2)}x_{i_2}$$

$$y^{(3)}=b_0^{(3)}+b_1^{(3)}x_{i_1}+b_2^{(3)}x_{i_2}+b_3^{(3)}x_{i_3}$$

$$\vdots$$

下面，我们讨论逐步消元变换的具体算法。

假若要从(5.4.1)中消去第 k 个未知量 $b_k(1\leqslant k\leqslant m)$，相当于在(5.4.1)的两边左乘一个变换矩阵

$$
\boldsymbol{D}_k=\begin{pmatrix}
1 & & & -\frac{a_{1k}}{a_{kk}} & & & \\
 & \ddots & & \vdots & & & \\
 & & 1 & -\frac{a_{k-1\,k}}{a_{kk}} & & & \\
 & & & \frac{1}{a_{kk}} & & & \\
 & & & -\frac{a_{k+1\,k}}{a_{kk}} & 1 & & \\
 & & & \vdots & & \ddots & \\
 & & & -\frac{a_{mk}}{a_{kk}} & & & 1
\end{pmatrix} \tag{5.4.2}
$$

即

$$
\boldsymbol{D}_k(a_{ij})(b_j)=\boldsymbol{D}_k(c_{ij})(l_{jy}) \tag{5.4.3}
$$

特别在对法方程(5.2.7)进行第一次消元变换时,有

$$
\begin{pmatrix}
a_{11}-\frac{a_{1k}a_{k1}}{a_{kk}} & \cdots & 0 & \cdots & a_{1m}-\frac{a_{1k}a_{km}}{a_{kk}} \\
\cdot & \cdots & \cdot & \cdots & \cdot \\
\frac{a_{k1}}{a_{kk}} & \cdots & 1 & \cdots & \frac{a_{km}}{a_{kk}} \\
\cdot & \cdots & \cdot & \cdots & \cdot \\
a_{m1}-\frac{a_{mk}a_{k1}}{a_{kk}} & \cdots & 0 & \cdots & a_{mm}-\frac{a_{mk}a_{km}}{a_{kk}}
\end{pmatrix}
\begin{pmatrix} b_1 \\ \vdots \\ b_k \\ \vdots \\ b_m \end{pmatrix}
$$

$$
=\begin{pmatrix}
1 & & -\frac{a_{1k}}{a_{kk}} & & \\
 & \ddots & \vdots & & \\
 & & \frac{1}{a_{kk}} & & \\
 & & \vdots & \ddots & \\
 & & -\frac{a_{mk}}{a_{kk}} & & 1
\end{pmatrix}
\begin{pmatrix} l_{1y} \\ \vdots \\ l_{ky} \\ \vdots \\ l_{my} \end{pmatrix} \tag{5.4.4}
$$

由此可以看出,消去一个未知量 b_k, 把预报因子 x_k 引入回归方程,就是用单位矩阵(δ_{ij})的第 k 个列向量,置换(a_{ij})中的相应列向量, (a_{ij})中的其它元素也进行相应的变换。这时,(c_{ij})中的第 k 个单位列向量,用新的列向量

$$
\left(-\frac{a_{1k}}{a_{kk}} \quad \cdots \quad -\frac{a_{k-1\,k}}{a_{kk}} \quad \frac{1}{a_{kk}} \quad -\frac{a_{k+1\,k}}{a_{kk}} \quad \cdots \quad -\frac{a_{mk}}{a_{kk}}\right)^{T}
$$

来代替,其它元素也作相应的变换。

经过若干步计算之后,回归方程中选入一些预报因子 x_k。这时, 在(a_{ij})中, 对应已选因子 x_k 的各列用相应的单位列向量置换,而(c_{ij})中的相应各列引入新的向量。所以,(a_{ij})中的单位列向量对应已选的预报因子,(c_{ij})中保留的单位列向量, 对应待选的预报因子, 总数m是不变的。因此,在数字计算机上进行计算时,就可以只用存放一个矩阵的 m^2 个存储单元,存放(a_{ij})和(c_{ij})中的非单位向量, 即在每一步计算中, 在形成单位列向量的地方, 存放新形

成的列向量

$$\left(-\frac{a_{1k}}{a_{kk}} \quad \cdots \quad \frac{1}{a_{kk}} \quad \cdots \quad -\frac{a_{mk}}{a_{kk}}\right)^T$$

把(a_{ij})、(c_{ij})中非单位列向量合并后的矩阵，仍用(a_{ij})表示。用a'_{ij}表示消元变换后新形成矩阵的元素，根据(5.4.2)、(5.4.3)，对两类消元变换过程，有着同样的消元算法，即

$$a'_{ij}=\begin{cases} a_{ij}-\dfrac{a_{ik}a_{kj}}{a_{kk}}, & i\neq k,\ j\neq k \\ \dfrac{a_{kj}}{a_{kk}}, & i=k,\ j\neq k \\ -\dfrac{a_{ik}}{a_{kk}}, & i\neq k,\ j=k \\ \dfrac{1}{a_{kk}}, & i=k,\ j=k \end{cases} \tag{5.4.5}$$

为了在矩阵消元计算过程中得到回归系数的估计值b_i，方便地确定已选和待选的预报因子，我们把m阶的系数矩阵(l_{ij})(5.2.8)，扩展成$(m+1)$阶矩阵：

$$\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} & l_{1y} \\ l_{21} & l_{22} & \cdots & l_{2m} & l_{2y} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ l_{m1} & l_{m2} & \cdots & l_{mm} & l_{my} \\ l_{y1} & l_{y2} & \cdots & l_{ym} & l_{yy} \end{pmatrix} \tag{5.4.6}$$

仍用(a_{ij})表示。这时，(a_{ij})中的最后一列，即第$(m+1)$列，不能选作消元的序数k。

对$(m+1)$阶矩阵(5.4.6)，不选$(m+1)$进行消元变换，用矩阵基本运算公式(5.4.5)进行消元计算时，易证：

(1) 当x_{i_1}尚待选入回归方程时，$a_{i_1y}a_{yi_1}>0$，贡献U_{i_1}为正；

(2) 当x_{i_2}已选入回归方程时，$a_{i_2y}a_{yi_2}<0$，贡献U_{i_2}为负；

(3) 计算结束时，在a_{i_2y}处得到回归系数β_{i_2}的估计值b_{i_2}，a_{yy}处给出残差平方和Q。这时，相应的a_{ij}上，给出矩阵(l^{ij})的计算结果。

综合舍选预报因子的统计检验(5.3.4)和矩阵的基本运算公式(5.4.5)，考虑到计算机对大量数据进行处理的特点，可把逐步回归计算的全过程分为三个基本步骤。

(一)计算相关矩阵

为简化公式的书写，记因变量

$$y_n=x_{ny}, \quad n=1,\ 2,\ \cdots,\ N$$

在回归分析计算中，(5.1.1)给出的观测数据共$N(m+1)$个，一般说来是很大的。为了提高计算结果的精度，减少计算机字长短带来的累积舍入误差的影响，我们用二次均值算法代替一次均值算法[7]，用规格化的相关矩阵(r_{ij})代替矩阵(5.4.6)。

计算均值

$$\bar{x}_i=\frac{1}{N}\sum_n x_{ni}+\frac{1}{N}\sum_{n_1}\left(x_{n_1i}-\frac{1}{N}\sum_n x_{ni}\right) \tag{5.4.7}$$

$$i=1,\ 2,\ \cdots,\ m,\ y$$

这里，$\frac{1}{N}\sum_{n_1}\left(x_{n_1i}-\frac{1}{N}\sum_n x_{ni}\right)$集中了一次均值$\frac{1}{N}\sum_n x_{ni}$计算过程中的主要舍入误差。基于同

样原因，用

$$l_{ij}=\sum_{n}(x_{ni}-\bar{x}_i)(x_{nj}-\bar{x}_j) \tag{5.4.8}$$

代替过去常用的计算格式 $\sum_{n}x_{ni}x_{nj}-N\bar{x}_i\bar{x}_j$，计算自乘、叉乘和。取

$$l_i=\sqrt{l_{ii}} \tag{5.4.9}$$

得相关系数

$$r_{ij}=\frac{l_{ij}}{l_i\cdot l_j} \tag{5.4.10}$$

显然，$r_{ii}=1$，$r_{ij}=r_{ji}$，且 $|r_{ij}|\leqslant 1$。因此，在计算机上，只需对 $i=1, 2, \cdots, m, y$；$j=1, \cdots, i$，进行计算，以节省计算量。

经(5.4.7)、(5.4.8)、(5.4.9)、(5.4.10)的计算，得到相应于(5.4.6)的$(m+1)$阶相关阵

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{2m} & r_{2y} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ r_{m1} & r_{m2} & \cdots & r_{mm} & r_{my} \\ r_{y1} & r_{y2} & \cdots & r_{ym} & r_{yy} \end{pmatrix} \tag{5.4.11}$$

用相关系数 r_{ij} 表示法方程(5.2.7)，有

$$\begin{cases} r_{11}b'_1+r_{12}b'_2+\cdots+r_{1m}b'_m=r_{1y} \\ r_{21}b'_1+r_{22}b'_2+\cdots+r_{2m}b'_m=r_{2y} \\ \cdots\cdots \\ r_{m1}b'_1+r_{m2}b'_2+\cdots+r_{mm}b'_m=r_{my} \end{cases} \tag{5.4.12}$$

其中

$$b'_i=\frac{l_i}{l_y}b_i \tag{5.4.13}$$

这时，残差平方和

$$Q=l_y^2-\sum_{i}b_il_{iy}=(1-\sum_{i}b'_ir_{iy})l_y^2$$

舍选预报因子的统计检验(5.3.4)

$$F_i=\frac{(N-m-1)b_i^2}{l^{ii}Q}=\frac{N-m-1}{1-\sum_{i}b'_ir_{iy}}\cdot\frac{(b'_i)^2}{r^{ii}}=V_i^2\frac{N-m-1}{1-\sum_{i}b'_ir_{iy}} \tag{5.4.14}$$

其中

$$V_i^2=\frac{(b'_i)^2}{r^{ii}}=\frac{U_i^2}{l_y^2} \tag{5.4.15}$$

是预报因子 x_i 的标准贡献。$r^{ii}=l_i^2l^{ii}$ 是m阶矩阵$(r^{ij})=(r_{ij})^{-1}$的主对角元素。

(二)因子舍选和消元变换

用(a_{ij})表示舍选计算过程中的$(m+1)$阶矩阵(5.4.11)。为了避免系数矩阵蜕化的问题，给出一个充分小的常数 $\varepsilon>0$，只对 $a_{ii}>\varepsilon$ 者$(i=1, 2, \cdots, m)$计算贡献

$$f_i=\frac{a_{iy}a_{yi}}{a_{ii}}$$

选取消元变换的主元。因子舍选和消元变换的过程是：

(1) 计算预报因子的贡献 V_i^2

在已选因子中，找出贡献最小的因子，即对 $f_i<0$ 者，找出 $V_{\min}^2=\min|f_i|$ 及其相应的已选预报因子 $x_{i_{\min}}$。显然，在开始计算时，不存在 $f_i<0$ 者。

在待选因子中，找出贡献最大的因子，即对 $f_i>0$ 者，找出 $V_{\max}^2=\max f_i$ 及其相应的待选预报因子 $x_{i_{\max}}$。

(2) 选取消元因子 x_k

当

$$\frac{\phi V_{\min}^2}{a_{yy}}<F_2 \tag{5.4.16}$$

时，已选预报因子 $x_{i_{\min}}$ 的贡献不再显著，在选入新的预报因子之前，应把它从回归方程中剔除出去。这时，取 $k=i_{\min}$，进行消元变换运算(5.4.5)。否则，检验

$$\frac{(\phi-1)V_{\max}^2}{a_{yy}-V_{\max}^2}<F_1 \tag{5.4.17}$$

是否成立。若(5.4.17)成立，这时，既无已选因子可舍，又无待选因子可选，可终止预报因子的舍选，转去计算回归结果(三)。若(5.4.17)不成立，待选预报因子 $x_{i_{\max}}$ 贡献显著，取 $k=i_{\max}$，通过消元变换(5.4.5)，把 $x_{i_{\max}}$ 选入回归方程。

这里，ϕ 为残差平方和 Q 的自由度。开始计算时，$\phi=N-1$。当回归方程增加一个预报因子时，自由度减少1，用 $\phi-1$ 代替 ϕ；当从回归方程中剔除一个因子时，自由度增加1，用 $\phi+1$ 代替 ϕ。F_1、F_2 是在显著水平 α 给定时，由 F-分布表中查到的两个常数。为了合理地选取预报因子，避免刚选入的因子立刻又被剔除出去的循环运算，一般取 $F_1>F_2$。

(3) 消元变换

对得到的 k 值，用(5.4.5)对矩阵 (a_{ij}) 进行基本变换，然后返回(1)、(2)，进行下一因子的舍选。

(三) 计算回归结果

当 $\phi<N-1$ 时，得到回归方程。根据(5.2.1)、(5.3.7)、(5.3.8)、(5.4.13)等，计算回归方程的检验值、回归系数及其预报值。

(1) 回归方程检验值

残差平方和 $\quad Q=l_y^2 a_{yy}$

标准差 $\quad \bar{\sigma}=l_y\sqrt{a_{yy}/\phi}$

复相关系数 $\quad R=[1-a_{yy}]^{1/2}$

F-检验值 $\quad F=\dfrac{\phi(1-a_{yy})}{(N-\phi-1)a_{yy}}$

(2) 回归系数及其标准差

当 $a_{iy}a_{yi}<0$ 时，预报因子 x_i 选入回归方程，可算出：

回归系数 $\quad b_i=\dfrac{l_y}{l_i}a_{iy}$

标准差 $\quad \bar{\sigma}_{b_i}=\dfrac{\bar{\sigma}}{l_i}\sqrt{a_{ii}}$

常数项 $\quad b_0=\bar{y}-\sum_i{}' b_i\bar{x}_i$

这里，$\sum_i{}'$ 表示只对已选的预报因子 x_i 求和。

(3) 计算预报值及其残差

预报值　　$y_n^* = b_0 + \sum_i' b_i x_{ni}$

残差　　$e_n = y_n - y_n^*$

$n=1, 2, \cdots, N$

根据逐步回归计算的三个基本步骤，可以得到图5.1所示的计算框图[1,9]。为了便于得到过渡回归方程及部分中间结果，把计算步骤(三)，适当分标在框图后面。

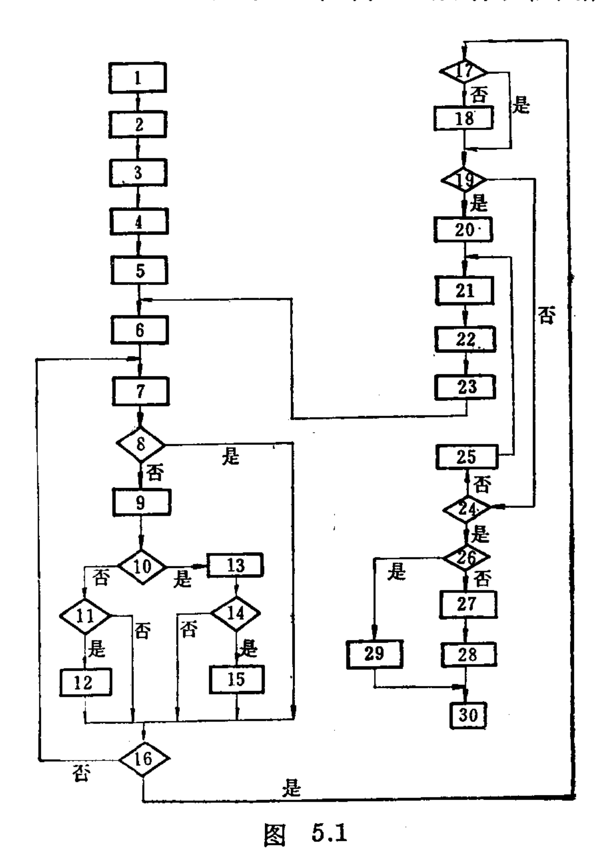


图 5.1

框 图 说 明

[框1] 输入原始观测数据和参量，其中包括：

(1) 数据参量

m 可能预报因子的个数

N 原始观测数据的组数

(2) 控制参量

F_1, F_2, ε 和输出控制信息

(3) 原始观测数据

$$(x_{n1}, x_{n2}, \cdots, x_{nm}; y_n)$$

$$n=1, 2, \cdots, N$$

[框2] 计算均值和叉乘和

$$\bar{x}_i = \frac{1}{N}\sum_n x_{ni}$$

$$l_{ij} = \sum_n (x_{ni} - \bar{x}_i)(x_{nj} - \bar{x}_j)$$

$$l_i=\sqrt{l_{ii}}$$

$$i=1, 2, \cdots, m, y \qquad j=1, 2, \cdots, i$$

这里,为简化符号,记 $x_{ny}=y_n$

[框 3] 计算下三角相关阵

$$r_{ij}=l_{ij}/l_il_j$$

$$i=1, 2, \cdots, m, y \qquad j=1, 2, \cdots, i$$

[框 4] 把下三角相关阵补为方阵

$$r_{ij}\Rightarrow a_{ij}\Rightarrow a_{ji}$$

[框 5] $N-1\Rightarrow\phi$ $l_y/\sqrt{N-1}\Rightarrow\bar{\sigma}$

[框 6] $0\Rightarrow i$

$10^4\Rightarrow V^2_{\min}$ $0\Rightarrow V^2_{\max}$

$0\Rightarrow i_{\min}$ $0\Rightarrow i_{\max}$

[框 7] $i+1\Rightarrow i$ $0\Rightarrow b_i$ $0\Rightarrow\bar{\sigma}_{b_i}$

[框 8] $a_{ii}<\varepsilon$?

检查矩阵的主对角元素,解决矩阵退化问题

[框 9] 计算第 i 个预报因子的贡献 f

$$a_{iy}a_{yi}/a_{ii}\Rightarrow f$$

[框 10] $f<0$?

根据 f 的符号,判别预报因子 x_i 是否已引入回归方程。

[框 11] $f>V^2_{\max}$?

[框 12] $f\Rightarrow V^2_{\max}$ $i\Rightarrow i_{\max}$

[框 13] 计算已引入变量的回归系数及其标准差

$$l_ya_{iy}/l_i\Rightarrow b_i \qquad \bar{\sigma}\sqrt{a_{ii}}/l_i\Rightarrow\bar{\sigma}_{b}$$

[框 14] $|f|<V^2_{\min}$?

[框 15] $|f|\Rightarrow V^2_{\min}$ $i=i_{\min}$

[框 16] $i=m$?

是否对 m 个预报因子检查完一遍?

[框 17] $\phi=N-1$?

[框 18] 计算回归系数的常数项

$$\bar{y}-\sum_i b_i\bar{x}_i\Rightarrow b_0$$

控制输出过渡回归方程的参量 b_i、$\bar{\sigma}_{bi}$

[框 19]

$$\frac{\phi V^2_{\min}}{a_{yy}}<F_2?$$

引入因子的显著性 F 检验

[框 20] 舍去已引因子 $x_{i\min}$

$$\phi+1\Rightarrow\phi \qquad i_{\min}\Rightarrow k$$

[框 21] 计算新矩阵的元素

$$a'_{ij}=\begin{cases} a_{ij}-\dfrac{a_{ik}a_{kj}}{a_{kk}} & i\neq k \quad j\neq k \\ a_{kj}/a_{kk} & i=k \quad j\neq k \\ -a_{ik}/a_{kk} & i\neq k \quad j=k \\ 1/a_{kk} & i=k \quad j=k \end{cases}$$

[框 22] 计算回归检验值

$$l_y^2 a_{yy} \Rightarrow Q \qquad l_y\sqrt{a_{yy}/\phi} \Rightarrow \bar{\sigma}$$

$$\sqrt{1-a_{yy}} \Rightarrow R \qquad \frac{\phi(1-a_{yy})}{(N-\phi-1)a_{yy}} \Rightarrow F$$

[框 23] 控制输出检验值

$$Q,\ \bar{\sigma},\ R,\ F$$

转向下轮计算

[框 24]

$$\frac{(\phi-1)V_{\max}^2}{a_{yy}-V_{\max}^2} < F_1?$$

未引入因子的显著性 F 检验

[框 25] 把因子 $x_{i_{\max}}$ 引入回归方程

$$\phi-1 \Rightarrow \phi \qquad i_{\max} \Rightarrow k$$

[框 26] $\phi=N-1$?

[框 27] 计算回归预报值及其残差

$$b_0+\sum_i b_i x_{ni} \Rightarrow y_n^*$$

$$y_n-y_n^* \Rightarrow e_n$$

[框 28] 输出回归结果

[框 29] 回归方程中选不进预报因子

[框 30] 停机

§5.5 逐步回归计算中的几个问题

最后，简略地讨论一下逐步回归分析计算中经常遇到的几个实际问题。

5.5.1 计算参量的选取

在逐步回归计算中，为避免系数矩阵蜕化，只对 $a_{ii}>\varepsilon$ 的因子计算贡献。一般计算中，可取 $0.0001<\varepsilon<0.001$。当取 $F_1=F_2=0$ 时，相当于用一般回归算法进行求解，选取所有能选的预报因子进入回归方程，这时可把 ε 取得更小一些。

在因子贡献显著性检验(5.4.16)、(5.4.17)中，临界值 F_1、F_2，是显著水平 α 和残差平方和自由度 $N-m-1$ 的函数。在逐步回归计算中，对给定的显著水平 α，F_1、F_2 随 m 而异，应取不同的值。在实际计算中，一般 $N\gg m$。为方便计，多把 F_1、F_2 取为常量，且 $F_1>F_2$。

在逐步回归计算中，结果回归方程中预报因子的个数 m，随 F_1、F_2 的取值不同而异(详见回归分析一例)。如果希望回归方程中多选进几个预报因子，可把选取预报因子的条件放宽一些，把 F_1、F_2 取的小一些；反之，F_1、F_2 要取得大一些。对 F_1、F_2，小者可在1左右，大则可取到10以上，一般多在2、4之间，视问题的物理性质而定。

5.5.2 回归效果的检验

在回归分析的实际应用中，对得到的回归方程，要进行预报可靠性和预报稳定性的检验。这里，只用回归计算中得到的一些统计检验值 Q、$\bar{\sigma}$、R、F，是不够的。在实际计算中，一般把观测数据(5.1.1)分为两部分，一部分(主要的)用来建立回归方程，一部分(少量的)

不参加回归方程的计算，用来检验回归的效果。当数据 N 太少时，不宜采用上述方法，可用蒙特卡洛方法进行模拟检验(参见本书“蒙特卡洛方法”一章及[6])，以确定回归效果。

为了提高回归方程预报的稳定性和可靠性，实际计算时，必须注意观测数据的组数 N 和预报因子的个数 m 之间的适当比例。一般要求 $N \gg m$，至少 N 应在 m 的 5 倍或 10 倍以上。

5.5.3 线性回归模型的推广

§5.4 给出的逐步回归算法，可以推广到更一般的情形[8,9,10]。

如取非线性回归方程

$$u=\exp[\beta_0+\beta_1 t_1+\beta_2 t_1^2+\beta_3 e^{t_1}+\cdots+\beta_m f(t_1,\ t_2,\ \cdots,\ t_k)+\varepsilon]$$

引进变量变换

$$\begin{aligned} y &= \ln u \\ x_1 &= t_1 \\ x_2 &= t_1^2 \\ x_3 &= e^{t_1} \\ &\vdots \\ x_m &= f(t_1,\ t_2,\ \cdots,\ t_k) \end{aligned}$$

得到线性回归模型(5.1.2)。这样，用线性回归算法，可拟合比较广泛的一类非线性回归方程。在曲线拟合，提取时间序列的趋势函数等要求自动选取最优拟合表达式的一类问题中，有着重要的应用[3,5]。

在回归模型(5.1.3)中，假定 ε_n 相互独立，服从 $N(0,\ \sigma^2)$ 分布。这种限制并非必要的。如[8]中，把回归模型(5.1.3)推广到更一般的情形，其中可假定误差 ε_n 是相关的。

5.5.4 逐步回归计算一例

下面，通过一个数值例子，更清楚地说明逐步回归的算法和步骤。

观测数据组数　$N=32$

可能预报因子个数　$m=4$

引入预报因子 F 检验临界值　$F_1=2+\pi/180$

剔除预报因子 F 检验临界值　$F_2=2$

控制参量　$\varepsilon=2^{-30}$

观测数据和计算结果由表 5.1 和表 5.2 给出。

结果中，计算误差未作修正。为方便计，有时用带幂的数值记法，如记

$$21.15=0.2115\times10^2$$

$$-0.034=-0.34\times10^{-1}$$

由相关矩阵 $(a_{ij}^{(0)})=(r_{ij})$，经过四步计算，既无已选因子可舍，又无预报因子可选时，终止回归计算，给出最后结果。

$$b_0=-5.11036,\quad b_1=0.506634,\quad b_4=0.501089$$

$$\bar{\sigma}_{b_1}=0.011987,\quad \bar{\sigma}_{b_4}=0.095482$$

表5.1 观测数据和预报值

编　号	自变量				因变量	预报值	残　差
	x_1	x_2	x_3	x_4	y	y^*	$e=y-y^*$
1	13	7	26	19	11.5	10.9966	0.5034
2	15	11	40	34	19.8	19.5262	0.2738
3	21	8	29	17	13.7	14.0475	−0.3475
4	19	12	15	33	21.6	21.0516	0.5484
5	27	11	13	27	22.3	22.0982	0.2018
6	32	10	21	15	19.1	18.6183	0.4817
7	17	8	18	16	11.7	11.5198	0.1802
8	26	10	35	23	19.4	19.5872	−0.1872
9	14	6	14	18	10.6	11.0021	−0.4021
10	28	13	21	34	25.5	26.1124	−0.6124
11	19	9	13	29	18.7	19.0473	−0.3473
12	12	10	19	38	19.3	20.0106	−0.7106
13	23	8	25	17	15.6	15.0607	0.5393
14	28	11	33	32	24.7	25.1102	−0.4102
15	21	9	18	19	15.3	15.0496	0.2504
16	35	14	24	34	29.8	29.6589	0.1411
17	16	6	19	14	10.2	10.0110	0.1890
18	24	10	32	26	19.8	20.0772	−0.2772
19	22	11	39	38	25.3	25.0770	0.2230
20	10	7	17	20	9.7	9.9778	−0.2778
21	18	8	34	22	14.8	15.0330	−0.2330
22	29	11	28	21	20.7	20.1049	0.5951
23	18	11	16	32	19.6	20.0439	−0.4439
24	16	10	15	34	20.3	20.0328	0.2672
25	18	7	23	14	11.1	11.0243	0.0757
26	23	11	29	29	20.7	21.0738	−0.3738
27	25	13	41	40	28.9	27.5991	1.3009
28	32	9	12	15	18.3	18.6183	−0.3183
29	36	11	37	18	21.5	22.1481	−0.6481
30	31	9	25	14	17.7	17.6105	0.0895
31	29	13	14	38	28.3	28.6234	−0.3234
32	18	10	11	35	21.6	21.5472	0.0528
最小值	10	6	11	14	9.7	9.9778	−0.7106
最大值	36	14	41	40	29.8	29.6589	1.3009
均　值	22.3437	9.8125	23.6250	25.4687	18.9718	18.9718	0
标准差①	6.8173	2.0377	8.9154	8.5586	5.4062	5.3883	0.4396

① 标准差 $s_i=\left[\frac{1}{32}\sum_{n}(x_{ni}-\bar{x}_i)^2\right]^{1/2}$

表 5.2 逐步回归计算过程中的矩阵(a_{ij})

$a_{ij}^{(0)}$ (r_{ij})	1.000000 0.567021 0.209841 -0.434669×10^{-1} 0.604392	0.567021 1.000000 0.207706 0.741491 0.956617	0.209841 0.207706 1.000000 0.100186 0.227809	-0.434669×10^{-1} 0.741491 0.100186 1.000000 0.765505	0.604392 0.956617 0.227809 0.765505 1.000000
f F	0.365289	0.915117 0.323429×10^{3} 引入x_2	0.518972×10^{-1}	0.585998	$\phi=31$
$a_{ij}^{(1)}$	0.678417 0.567021 0.920669×10^{-1} -0.463908 0.619696×10^{-1}	-0.567021 1.000000 -0.207706 -0.741491 -0.956617	0.920669×10^{-1} 0.207706 0.956858 -0.538259×10^{-1} 0.291140×10^{-1}	-0.463908 0.741491 -0.538259×10^{-1} 0.450190 0.561817×10^{-1}	0.619696×10^{-1} 0.956617 0.291140×10^{-1} 0.561817×10^{-1} 0.848825×10^{-1}
f F	0.566×10^{-2}	-0.915117 0.323429×10^{3}	0.885844×10^{-3}	0.701122×10^{-2} 0.261104×10^{1} 引入x_4	$\phi=30$
$a_{ij}^{(2)}$	0.200443 0.133110×10^{1} 0.366009×10^{1} -0.103047×10^{1} 0.119863	-0.133110×10^{1} 0.222128×10^{1} -0.296361 -0.164706×10^{1} -0.864083	0.366009×10^{-1} 0.296361 0.950422 -0.119562 0.358312×10^{-1}	0.103047×10^{1} -0.164706×10^{1} 0.119562 0.222128×10^{1} -0.124795	0.119863 0.864083 0.358312×10^{-1} 0.124795 0.778713×10^{-1}
f F	0.716770×10^{-1} 0.323999×10^{3} 引入x_1	-0.336130	0.135085×10^{-2}	-0.701122×10^{-2} 0.261104×10^{1}	$\phi=29$
$a_{ij}^{(3)}$	0.498893×10^{1} -0.664079×10^{1} -0.182599 0.514094×10^{1} -0.597989	-0.664079×10^{1} 0.110609×10^{2} -0.533015×10^{-1} -0.849020×10^{1} -0.680957×10^{-1}	0.182599 0.533015×10^{-1} 0.943739 0.686008×10^{-1} 0.139443×10^{-1}	0.514094×10^{1} -0.849020×10^{1} -0.686008×10^{-1} 0.751887×10^{1} -0.741006	0.597989 0.680954×10^{-1} 0.139443×10^{-1} 0.741006 0.619432×10^{-2}
f F	-0.716770×10^{-1}	-0.419225×10^{-3} 0.189501×10^{1} 剔除x_2	0.206035×10^{-3}	-0.730282×10^{-1}	$\phi=28$
$a_{ij}^{(4)}$	0.100189×10^{1} -0.600385 0.214601 0.435493×10^{-1} -0.638873	0.600385 0.904087×10^{-1} 0.481892×10^{-2} 0.767588 0.615644×10^{-2}	0.214601 0.481892×10^{-2} 0.943996 0.109514 0.142724×10^{-1}	0.435493×10^{-1} -0.767588 -0.109514 0.100189×10^{1} -0.793275	0.638873 0.615644×10^{-2} 0.142724×10^{-1} 0.793275 0.661354×10^{-2}
f F	-0.407387 0.178600×10^{4}	0.419225×10^{-3} 0.189501×10^{1}	0.215787×10^{-3}	-0.628096	$\phi=29$

得回归方程

$$y^* = -5.11036 + 0.506634x_1 + 0.501089x_4$$

估计值 y_n^* 及残差 e_n，由表5.1给出。

回归检验值:

$$Q = 6.1854,\ \bar{\sigma} = 0.461833,\ R = 0.996687,\ F = 2177.97$$

从上例不难看出，对不同的 F 检验临界值，可以得到不同的回归方程。

如取 $F_1 = \pi/180$、$F_2 = 0$ 时，得到回归方程

$$y^* = -5.44484 + 0.47207x_1 + 0.17857x_2 + 0.00896x_3 + 0.46743x_4$$

包括全部四个预报因子，复相关 $R = 0.997001$。当取 $F_1 = 3 + \pi/180$, $F_2 = 3$ 时，只能选入一个预报因子，回归方程为

$$y^* = -5.93182 + 2.53795x_2$$

这时，复相关系数 $R = 0.956617$。

参 考 资 料

[1] A. 拉尔斯登, H. S. 维尔夫著，徐献瑜等译，《数字计算机上用的数学方法》，§17 多重回归分析，pp. 302-316, 科学技术出版社，1963。

[2] 森口繁一著，刘璋温译，《统计分析》，科学技术出版社，1961。

[3] Anderss R. S., Osborne M. R. (eds.), "Least Squares Methods in Data Analysis", Australian National Univesity Computer Centre Publ., 1969.

[4] Aubert E. J., Lund I. A., Thomasell A., "Some objective six-hour prediction prepared by statistical methods", J. of meteor., **16**(4), 1959, pp. 436-446.

[5] Draper N. R., Smith H., "Seleting the "best" regression equation", «Computer Application in the Earth Science: Colloquium on Trend Analysis», 1967, pp. 1-9.

[6] Lund I. A., "A Monte Carlo method for testing the statistical significance of a regression equation", J. of Appl. Meteor., **9**(3), 1970, pp. 330-332.

[7] Neely P. M., "Comparision of several algorithms for computation means, standard deviation and correlation coefficients", Communication of the ACM., **9**(7), 1966, pp. 496-499.

[8] Rao C. R., "Linear Statistical Inference and Its application", John Wiley 1965.

[9] Smillie K. W., "An Introduction to Regression and Correlation", 1966.

[10] Williams E. J., "Regression Analysis", John Wiley, 1959.

第六章　时间序列分析

§6.1　时 间 序 列

对生产斗争、科学实验过程中的一个变量或一组变量 $x(t)$ 进行测量，在时刻 $t_1<t_2<\cdots<t_N$，得到以时间 t 为参量的有序数集合

$$x(t_1),\ x(t_2),\ \cdots,\ x(t_N)$$

一般把它叫做时间序列。这里，为方便计，把参量 t 叫作时间。在实际问题中，参量 t 可具有不同的物理意义。

在生产斗争和科学实验过程中，出现时间序列的例子是很多的。如描述太阳黑子活动的黑子相对数，一地区的年降水量，一河流的月最高水位，一地震带上的地震目录，一动力学系统的输出、输入等，都是时间序列的实际例子。我们把用来处理这种时间序列的方法，称为时间序列分析。

为简单计，下面只讨论在等距时间间隔 h 上取值的一维和多维时间序列：

$$x_1,\ x_2,\ \cdots,\ x_n,\ \cdots,\ x_N \tag{6.1.1}$$

这里

$$x_n=x(t_n)=x(t_0+nh)$$
$$n=1,\ 2,\ \cdots,\ N$$

其中 t_0 是量测的开始时间。在今后的分析计算中，t_0 和 h 只有相对的意义，一般并不在计算公式中出现。对一个多维时间序列，如一个三维时间序列，有

$$x_n=\begin{pmatrix} x_1(t_n) \\ x_2(t_n) \\ x_3(t_n) \end{pmatrix}=(x_1(t_n),\ x_2(t_n),\ x_3(t_n))^T$$

在许多实际问题中，常常需要根据 $x(t)$ 在 $t\leqslant t_N$ 时得到的测量数据(6.1.1)，分析时间序列 $x(t)$ 的规律，推断产生时间序列 $x(t)$ 的物理系统的性质，预报 $t>t_N$ 时 $x(t)$ 的取值。

下面，举一个简单例子来作说明。

给出某地区100年间的诸年年降水量 $x(t)$，组成一个 $N=100$ 的一维时间序列。我们希望通过对这组数据的分析，研究一下该地区的旱涝规律，同时能对该地区今后几年的降水量有所预报。

当 $x(t)$ 是一个多维时间序列时，如取

$$x(t)=(x_1(t),\ x_2(t),\ x_3(t))^T$$

其中，$x_1(t)$ 表示某地区的年降水量；$x_2(t)$ 表示该地区同年的冬季平均气温；$x_3(t)$ 表示同年的太阳黑子相对数。我们希望通过对三维时间序列 $x(t)$ 的分析，研究一下该地区的降水和冬季平均气温、太阳黑子活动之间可能存在的关系，为降水预报服务。

无疑，上述一类分析、预报问题，应用广泛，有着一定的实际意义。

一个时间序列 $x(t)$，如果它的取值由一个完全确定的数学函数给出时，如

$$x(t)=a+b\cos\omega t \tag{6.1.2}$$

其中，a、b、ω 为给定的常量，称为确定的。实际问题中的时间序列是复杂的。一般说来，$x(t)$不可能像(6.1.2)那样，用一个完全确定的数学函数给出来，但可以用一个概率分布函数组给出 $x(t)$未来取值状况的统计描述。我们称这种类型的时间序列为随机的。这里讨论的就是随机时间序列的统计分析和统计预报问题。今后，把随机时间序列简称为时间序列。

为了处理随机时间序列的分析和预报问题，在数学上，就是把给出的测量数据(6.1.1)看作随机过程 $X(t)$的一个现实，一个样本函数。通过对现实(6.1.1)的分析，估计随机过程 $X(t)$的总体特性，预测 $X(t)$未来取值的概率分布，从而给出 $t>t_N$ 时 $X(t)$的预报值

$$x_{N+1}^*,\ x_{N+2}^*,\ \cdots,\ x_{N+L}^*$$

要求它们和 $X(t)$未来实测值

$$x_{N+1},\ x_{N+2},\ \cdots,\ x_{N+L}$$

之差

$$\delta_l=x_{N+l}-x_{N+l}^* \quad (l=1,\ 2,\ \cdots,\ L)$$

的平方期望

$$E[\delta_l^2]=E[(x_{N+l}-x_{N+l}^*)^2]$$

最小。

在实际问题中，可以通过不同的方法，如等距采样法、累积法、特征值法等，给出离散数字化的等距时间序列(6.1.1)。

下面，我们根据时间序列(6.1.1)的统计性质，构造不同的数学概率模型，讨论 $x(t)$的分析和预报问题。

§6.2 平稳时间序列分析

由随机过程 $X(t)$给出的时间序列(6.1.1)是非常一般的。为便于讨论，我们先从一类简单、常用的平稳时间序列入手，讨论一维和多维平稳时间序列的分析和预报问题。然后，再转向一般时间序列的分析和预报。

6.2.1 一维平稳时间序列分析

一个随机过程 $X(t)$，如果它的统计性质不随时间原点的推移而变化，称为平稳的[2,3,4]。由平稳随机过程 $X(t)$的一个现实函数 $x(t)$，经离散数字化处理后，得到平稳时间序列(6.1.1)。根据实际问题的性质和要求，这里只讨论广义平稳随机过程和广义平稳时间序列的统计分析和统计预报问题，并把广义平稳随机过程简称为平稳过程。

一个平稳的随机过程 $X(t)$，具有两个显著的特点：(1)它的量测数据 $x(t)$围绕在一个固定不变的水平线附近，均匀地随机摆动；(2)在任意两个不同时刻 t 和 $t+\tau$ 上得到的随机数据 $X(t)$和 $X(t+\tau)$之间的统计性质只是它们时间间隔 τ 的函数，不依赖于时间原点 t_0 的位置。在数学上，这意味着随机过程 $X(t)$的数学期望和方差

$$\left.\begin{aligned}\mu(t)&=E[X(t)]=\mu\\ \sigma^2(t)&=E[(X(t)-\mu)^2]=\sigma^2\end{aligned}\right\} \tag{6.2.1}$$

取常量，相关函数

$$\begin{aligned}\rho(t,\ t+\tau)&=E\left[\left(\frac{X(t)-\mu(t)}{\sigma(t)}\right)\left(\frac{X(t+\tau)-\mu(t+\tau)}{\sigma(t+\tau)}\right)\right]\\&=E\left[\left(\frac{X(t)-\mu}{\sigma}\right)\left(\frac{X(t+\tau)-\mu}{\sigma}\right)\right]\\&=E(\tilde{X}(t)\cdot\tilde{X}(t+\tau))=\rho(\tau)\end{aligned}\tag{6.2.2}$$

与时间 t 无关，只是它们时间间隔 τ 的函数。这里

$$\tilde{X}(t)=\frac{X(t)-\mu}{\sigma}\tag{6.2.3}$$

表示经过标准化处理后，均值为 0，方差为 1 的平稳过程。

根据相关函数 $\rho(\tau)$ 的定义(6.2.2)，有

$$\rho(\tau)=\rho(-\tau)\tag{6.2.4}$$

是 τ 的偶函数，且

$$|\rho(\tau)|\leqslant\rho(0)=1$$

在 $\tau=0,\ 1,\ 2,\ \cdots,\ m-1$ 时，$\rho(\tau)$ 的值组成一个 m 阶的相关矩阵

$$(\rho(i-j))=\begin{pmatrix}\rho(0)&\rho(1)&\rho(2)&\cdots&\rho(m-1)\\\rho(1)&\rho(0)&\rho(1)&\cdots&\rho(m-2)\\\rho(2)&\rho(1)&\rho(0)&\cdots&\rho(m-3)\\\cdot&\cdot&\cdot&\cdots&\cdot\\\rho(m-1)&\rho(m-2)&\rho(m-3)&\cdots&\rho(0)\end{pmatrix}$$

是正定实对称的，称为 Toeplitz 矩阵[8]。

假定 $X(t)$ 是一个各态历经的平稳过程[4]，可以用时间平均代替总体平均。根据平稳随机过程 $X(t)$ 的一个现实序列(6.1.1)，可以给出它的数学期望 μ、方差 σ^2 和相关函数 $\rho(\tau)$ 的无偏或渐近无偏的统计估计值：

$$\left.\begin{aligned}\bar{x}&=\frac{1}{N}\sum_{n=1}^{N}x_n\\s^2&=\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})^2\\r(\tau)&=\frac{1}{N-\tau}\sum_{n=1}^{N-\tau}\left(\frac{x_n-\bar{x}}{s}\right)\left(\frac{x_{n+\tau}-\bar{x}}{s}\right)\\&=\frac{1}{N-\tau}\sum_{n=1}^{N-\tau}\tilde{x}_n\cdot\tilde{x}_{n+\tau}\qquad\tau=0,\ 1,\ 2,\ \cdots,\ m,\quad m<N\end{aligned}\right\}\tag{6.2.5}$$

在实际计算中，为了给出相关函数 $\rho(\tau)$ 的有效统计估值 $r(\tau)$，时间序列(6.1.1)的采样时间间隔 h 不宜过大，采样点数 N 不能太小。利用(6.2.5)进行计算时，一般要求 $N>50$，m 不应过大，如可取 $m<\frac{N}{4}$，常用的 m 值约在 $N/10$ 左右。

相关函数 $\rho(\tau)$，在时域上描述了平稳随机过程 $X(t)$ 的基本统计特性，它的傅里叶变换

$$g(\omega)=\int_{-\infty}^{+\infty}\rho(\tau)e^{-j\omega\tau}d\tau=2\int_{0}^{+\infty}\rho(\tau)\cos\omega\tau d\tau\tag{6.2.6}$$

从频域上描述了平稳过程 $X(t)$ 的基本统计特性，称为平稳过程的功率谱密度[2,4,7,10]。在动力学系统研究中，谱密度 $g(\omega)$ 有着重要的应用。

用(6.2.5)给出的估计值 $r(\tau)$，代替(6.2.6)中的理论相关函数 $\rho(\tau)$，可以得到谱密度 $g(\omega)$ 的渐近无偏估计值[1,4,7]：

$$g^*(k)=1+2\sum_{\tau=1}^{m-1}r(\tau)\cos\frac{k\tau\pi}{m}+r(m)\cos k\pi\quad k=0,\ 1,\ \cdots,\ m \tag{6.2.7}$$

在实际计算中，为了减少采样误差，可以利用三点平滑公式

$$\left.\begin{aligned}&\bar{g}_0=\frac{1}{2}[g^*(0)+g^*(1)]\\&\bar{g}_k=\frac{1}{4}[g^*(k-1)+2g^*(k)+g^*(k+1)]\quad k=1,\ 2,\ \cdots,\ m-1\\&\bar{g}_m=\frac{1}{2}[g^*(m-1)+g^*(m)]\end{aligned}\right\} \tag{6.2.8}$$

给出谱密度 $g(\omega)$ 的更优估计值。

为了能对平稳时间序列(6.1.1)进行预报，我们引入m阶自回归预报模型：

$$\tilde{x}_t=\phi_1\tilde{x}_{t-1}+\phi_2\tilde{x}_{t-2}+\cdots+\phi_m\tilde{x}_{t-m}+\varepsilon_t=\sum_{j=1}^{m}\phi_j\tilde{x}_{t-j}+\varepsilon_t \tag{6.2.9}$$

这里，自回归系数 $\phi_1,\ \phi_2,\ \cdots,\ \phi_m$ 为常量。在我们的预报问题中，m 和 $\phi_j(j=1,\ 2,\ \cdots,\ m)$ 为待定的参量。ε_t 是均值为零、方差为 σ_ε^2 的白噪声[4,5]。它的相关函数

$$\rho_\varepsilon(\tau)=\begin{cases}1, & 当\ \tau=0\ 时\\0, & 当\ \tau\neq0\ 时\end{cases}$$

且当 $i>0$ 时，

$$E[\tilde{x}_{t-i}\varepsilon_t]=0 \tag{6.2.10}$$

白噪声 ε_t 表示测量过程中存在的各种相互独立的随机干扰和今后预报中出现的误差。

用 $\tilde{x}_{t-i}$ 乘(6.2.9)的两边，得到方程

$$\tilde{x}_t\tilde{x}_{t-i}=\sum_{j=1}^{m}\phi_j\tilde{x}_{t-j}\tilde{x}_{t-i}+\varepsilon_t\tilde{x}_{t-i} \tag{6.2.11}$$

根据(6.2.2)、(6.2.10)，对(6.2.11)的两边取数学期望，得到自回归系数 ϕ_j 满足的 m 阶差分方程

$$\rho(i)=\sum_{j=1}^{m}\rho(i-j)\phi_j$$

这里 $i>0$。取 $i=1,\ 2,\ \cdots,\ m$，得到 ϕ_j 满足的m阶线性方程组

$$\left.\begin{aligned}&\phi_1+\rho(1)\phi_2+\cdots+\rho(m-1)\phi_m=\rho(1)\\&\rho(1)\phi_1+\phi_2+\cdots+\rho(m-2)\phi_m=\rho(2)\\&\qquad\cdots\cdots\\&\rho(m-1)\phi_1+\rho(m-2)\phi_2+\cdots+\phi_m=\rho(m)\end{aligned}\right\} \tag{6.2.12}$$

如果用相关函数 $\rho(\tau)$ 的估计值 $r(\tau)$ 代替(6.2.12)中的 $\rho(\tau)$，得到自回归系数 ϕ_j 的估计值 b_j 满足的m阶线性方程组。用m阶 Toeplitz 矩阵$(r(i-j))$表示时，有

$$(r(i-j))\begin{pmatrix}b_1\\b_2\\\vdots\\b_m\end{pmatrix}=\begin{pmatrix}r(1)\\r(2)\\\vdots\\r(m)\end{pmatrix} \tag{6.2.13}$$

显然，可用不同的方法求解方程组(6.2.13)，其中比较好的有用逐步回归算法的直接求解和利用 Toeplitz 矩阵$(r(i-j))$特有性质的递推求解。

用逐步回归算法求解(6.2.13)，对给定的舍选预报因子 F-检验的临界值 F_1、F_2，对自

回归模型(6.2.9)的因子$\tilde{x}_{t-j}(j=1, 2, \cdots, m)$进行舍选，达到去粗取精，去伪存真的目的，同时又可避免系数矩阵$(r(i-j))$出现病态或蜕化时的计算处理问题。根据(6.1.1)中给出的测量数据个数N，确定可能计算相关系数的m值，考虑到预报问题的物理性质和所用计算机内存容量的限制，选取自回归预报方程(6.2.9)的适当预报阶数m，用逐步回归算法，在计算过程中，自动挑选实际预报时的阶数。关于逐步回归算法，在本书《回归分析》一章及参考资料[1]中，都有详细的叙述。

下面，我们以$m=2$、3为例，说明如何利用Toeplitz矩阵的性质，递推求解(6.2.13)。为此，记m阶线性方程组(6.2.13)的解b_j为$b_{mj}(j=1, 2, \cdots, m)$。在$m=2$、3时，有

$$\begin{cases} b_{21}+r(1)b_{22}=r(1) \\ r(1)b_{21}+b_{22}=r(2) \end{cases} \tag{6.2.14}$$

和

$$\begin{cases} b_{31}+r(1)b_{32}+r(2)b_{33}=r(1) \\ r(1)b_{31}+b_{32}+r(1)b_{33}=r(2) \\ r(2)b_{31}+r(1)b_{32}+b_{33}=r(3) \end{cases} \tag{6.2.15}$$

在(6.2.15)中，前两个方程可改写为

$$\begin{cases} b_{31}+r(1)b_{32}=r(1)-b_{33}r(2) \\ r(1)b_{31}+b_{32}=r(2)-b_{33}r(1) \end{cases}$$

根据(6.2.14)，

$$\begin{pmatrix} b_{21} \\ b_{22} \end{pmatrix}=\begin{pmatrix} 1 & r(1) \\ r(1) & 1 \end{pmatrix}^{-1}\begin{pmatrix} r(1) \\ r(2) \end{pmatrix}$$

可得

$$\begin{pmatrix} b_{31} \\ b_{32} \end{pmatrix}=\begin{pmatrix} b_{21} \\ b_{22} \end{pmatrix}-b_{33}\begin{pmatrix} b_{22} \\ b_{21} \end{pmatrix}$$

即

$$\begin{cases} b_{31}=b_{21}-b_{33}b_{22} \\ b_{32}=b_{22}-b_{33}b_{21} \end{cases} \tag{6.2.16}$$

将(6.2.16)代入(6.2.15)的最后一个方程，整理后得

$$b_{33}=\frac{r(3)-b_{21}r(2)-b_{22}r(1)}{1-b_{21}r(1)-b_{22}r(2)} \tag{6.2.17}$$

因此，算出b_{21}、b_{22}后，根据(6.2.17)、(6.2.16)，可递推算出三阶自回归系数的估计值b_{33}、b_{31}、b_{32}。

假若已经得到m阶线性方程组(6.2.13)的解$b_{m1}, b_{m2}, \cdots, b_{mm}$，利用上述方法，不难得到$(m+1)$阶方程的解

$$\begin{cases} b_{m+1\,m+1}=\dfrac{r(m+1)-\sum\limits_{j=1}^{m}b_{mj}r(m+1-j)}{1-\sum\limits_{j=1}^{m}b_{mj}r(j)} \\ b_{m+1\,j}=b_{mj}-b_{m+1\,m+1}\,b_{m\,m+1-j} \quad j=1, 2, \cdots, m \end{cases} \tag{6.2.18}$$

用递推算法(6.2.18)求解方程组(6.2.13)，乘加运算量正比于m^2，和直接算法求解(6.2.13)的运算量m^3相比，计算速度提高了一个数量级。特别，这一算法只需要$3m$个存储单元，在计算机上，可以求解更高阶的自回归模型(6.2.9)。

由(6.2.13)解出b_j后，根据(6.2.9)、(6.2.3)，得到x_{N+1}的预报值

$$x_{N+1}^{*}=\bar{x}+s\sum_{j=1}^{m}b_{j}\tilde{x}_{N+1-j} \tag{6.2.19}$$

类推，可给出预报值 x_{N+2}^{*}, …, x_{N+2}^{*}。

6.2.2 多维平稳时间序列分析

这里，我们以三维时间序列为例，讨论多维平稳时间序列的分析和预报问题。向更高维推广时，完全是类似的[6]。

设
$$X(t)=(X_1(t),X_2(t),X_3(t))^T$$
是一个三维平稳随机过程。在数学上，表示随机过程 $X_1(t)$、$X_2(t)$、$X_3(t)$ 的数学期望和方差

$$\left.\begin{aligned}\mu_i(t)&=E[X_i(t)]=\mu_i\\ \sigma_i^2(t)&=E[(X_i(t)-\mu_i)^2]=\sigma_i^2\end{aligned}\right\} \tag{6.2.20}$$

取常量，它们的自相关函数

$$\rho_{ii}(\tau)=E(\tilde{X}_i(t)\tilde{X}_i(t+\tau)) \tag{6.2.21}$$

和互相关函数

$$\rho_{ij}(\tau)=E(\tilde{X}_i(t)\tilde{X}_j(t+\tau)) \tag{6.2.22}$$

与 t 无关，只是它们的时间间隔 τ 的函数。

根据相关函数的定义(6.2.21)、(6.2.22)，可知

$$\rho_{ii}(\tau)=\rho_{ii}(-\tau),\quad \rho_{ij}(\tau)=\rho_{ji}(-\tau) \tag{6.2.23}$$

因此，多维平稳随机过程

$$\tilde{X}(t)=\left(\frac{X_1(t)-\mu_1}{\sigma_1},\ \frac{X_2(t)-\mu_2}{\sigma_2},\ \frac{X_3(t)-\mu_3}{\sigma_3}\right)^T$$

的相关函数矩阵

$$R(\tau)=E[\tilde{X}(t)\tilde{X}^T(t+\tau)]=\begin{pmatrix}\rho_{11}(\tau)&\rho_{12}(\tau)&\rho_{13}(\tau)\\ \rho_{21}(\tau)&\rho_{22}(\tau)&\rho_{23}(\tau)\\ \rho_{31}(\tau)&\rho_{32}(\tau)&\rho_{33}(\tau)\end{pmatrix} \tag{6.2.24}$$

根据(6.2.23)、(6.2.24)，可知

$$R^T(\tau)=R(-\tau) \tag{6.2.25}$$

若 $X(t)$ 是一个各态历经的多维平稳过程，根据它的一个现实序列(6.1.1)，可以给出它的数学期望 μ_i，方差 σ_i^2 和自、互相关函数 $\rho_{ii}(\tau)$、$\rho_{ij}(\tau)$ 的无偏或渐近无偏的统计估计值：

$$\left.\begin{aligned}\bar{x}_i&=\frac{1}{N}\sum_{n=1}^{N}x_i(n)\\ s_i^2&=\frac{1}{N}\sum_{n=1}^{N}(x_i(n)-\bar{x}_i)^2\\ r_{ij}(\tau)&=\frac{1}{N-\tau}\sum_{n=1}^{N-\tau}\left(\frac{x_i(n)-\bar{x}_i}{\sigma_i}\right)\left(\frac{x_j(n+\tau)-\bar{x}_j}{\sigma_j}\right)\\ &=\frac{1}{N-\tau}\sum_{n=1}^{N-\tau}\tilde{x}_i(n)\tilde{x}_j(n+\tau)\\ i,\ j&=1,\ 2,\ 3;\quad \tau=0,\ 1,\ \cdots,\ m\end{aligned}\right\} \tag{6.2.26}$$

为了预报多维平稳时间序列，取 m 阶自回归预报模型

$$\begin{pmatrix}\tilde{x}_1(t)\\ \tilde{x}_2(t)\\ \tilde{x}_3(t)\end{pmatrix}=\sum_{j=1}^{m}\begin{pmatrix}\phi_{11}^{(j)} & \phi_{12}^{(j)} & \phi_{13}^{(j)}\\ \phi_{21}^{(j)} & \phi_{22}^{(j)} & \phi_{23}^{(j)}\\ \phi_{31}^{(j)} & \phi_{32}^{(j)} & \phi_{33}^{(j)}\end{pmatrix}\begin{pmatrix}\tilde{x}_1(t-j)\\ \tilde{x}_2(t-j)\\ \tilde{x}_3(t-j)\end{pmatrix}+\begin{pmatrix}\varepsilon_1(t)\\ \varepsilon_2(t)\\ \varepsilon_3(t)\end{pmatrix}$$

或简记为

$$\tilde{x}(t)=\sum_{j=1}^{m}\Phi_j\tilde{x}(t-j)+\varepsilon(t) \tag{6.2.27}$$

这里，$\tilde{x}(t)$是均值为0、方差为1的三维平稳随机过程。白噪声$\varepsilon(t)=(\varepsilon_1(t), \varepsilon_2(t), \varepsilon_3(t))^T$，表示测量过程中存在的随机干扰。回归系数矩阵

$$\Phi_j=\begin{pmatrix}\phi_{11}^{(j)} & \phi_{12}^{(j)} & \phi_{13}^{(j)}\\ \phi_{21}^{(j)} & \phi_{22}^{(j)} & \phi_{23}^{(j)}\\ \phi_{31}^{(j)} & \phi_{32}^{(j)} & \phi_{33}^{(j)}\end{pmatrix}$$

为进行预报时的待定参量。

用$\tilde{x}^T(t-i)$右乘(6.2.27)的两边，取数学期望，得到回归系数矩阵Φ_j满足的m阶差分方程

$$R(-i)=\sum_{j=1}^{m}\Phi_jR(j-i) \tag{6.2.28}$$

将(6.2.28)的两端转置，根据(6.2.25)，得Φ_j满足的三个$3m$阶线性方程组

$$R(i)=\sum_{j=1}^{m}R(i-j)\Phi_j^T \quad i=1,\ 2,\ \cdots,\ m \tag{6.2.29}$$

把由(6.2.26)算得的自、互相关函数$r_{ij}(\tau)$代入(6.2.29)，求解回归系数矩阵Φ_j满足(6.2.29)的估计矩阵B_j，给出$\tilde{x}(N+1)$的预报值

$$\tilde{x}^*(N+1)=\sum_{j=1}^{m}B_j\tilde{x}(N+1-j) \tag{6.2.30}$$

对一个K维平稳时间序列，用自回归模型(6.2.27)进行预报时，为了解出系数矩阵Φ_j的估计值B_j，需要求解k个km阶的线性方程组。显然，当k、m取值比较大时，计算是困难的。

§6.3 时间序列的平稳性检验

在$X(t)$是各态历经平稳随机过程的假定下，利用m阶自回归预报模型(6.2.9)、(6.2.27)，得到自回归系数ϕ_j、Φ_j满足的线性方程组(6.2.13)、(6.2.29)。当在实际问题中，利用ϕ_j、Φ_j的估计值b_j、B_j进行预报时，对一些问题是成功的，有着一定的预报参考价值，而对另外一些问题，预报是不成功的，甚至是很不成功的。显然，对一些实际问题，分析预报成败的原因是很复杂的，而为了达到不断提高预报效果的目的，这种分析又是必不可少的。下面，我们讨论一些数学方法，帮助我们分析时间序列模型预报成败的原因。

大家熟知这样一个事实：当我们研究一类物理现象时，总是试图建立各种不同类型的数学概率模型，帮助我们描述和分析问题。在§6.1中给出的降水平稳时间序列分析、预报的例子里，就是利用平稳随机模型来帮助我们描述大气物理中的一些现象的。很显然，利用上述模型进行预报成功与否，在很大程度上，取决于大气物理中的这类现象是否具有模型中要求的统计平稳性。

一般来讲，一个物理随机过程，如果它的实验条件基本保持不变的话，可以看作是平稳

的。但是，对一个实际问题，这一判据既不确切，也不易应用。因此，在实际问题中，多借助于统计分析方法，通过对一组测量数据(6.1.1)的统计分析，检验平稳性的假设是否合理。

如前所述，一个平稳的随机过程 $X(t)$，具有两个基本特点，即它的数学期望、方差取常值，相关函数只是时间间隔 τ 的函数，不依赖于时间 t。在自回归模型的建立和计算过程中，也正是利用了这两个性质。因此，时间序列的平稳性检验问题，也就是通过 $X(t)$ 的一个现实序列(6.1.1)，检验随机过程 $X(t)$ 是否具有这两个性质。

假若随机过程 $X(t)$ 的现实序列(6.1.1)足够长，即 N 取值足够大，如取 $N=kM$。按长度 M 把现实(6.1.1)分为 k 段子序列：

$$x_{i1},\ x_{i2},\ \cdots,\ x_{iM}$$

这里

$$x_{ij}=x_{(i-1)M+j}\quad i=1,\ 2,\ \cdots,\ k;\ j=1,\ 2,\ \cdots,\ M$$

根据公式(6.2.5)，可以计算各个子序列的均值、方差和相关函数的估计值：

$$\left.\begin{aligned}\bar{x}_i&=\frac{1}{M}\sum_{j=1}^{M}x_{ij}\\ s_i^2&=\frac{1}{M}\sum_{j=1}^{M}(x_{ij}-\bar{x}_i)^2\\ r_i(\tau)&=\frac{1}{M-\tau}\sum_{j=1}^{M-\tau}\left(\frac{x_{ij}-\bar{x}_i}{s_i}\right)\left(\frac{x_{ij+\tau}-\bar{x}_i}{s_i}\right)\end{aligned}\right\}\tag{6.3.1}$$

根据平稳性假设，利用(6.3.1)得到各个子序列的样本均值 $\bar{x}_i$，方差 s_i^2 和相关函数 $r_i(\tau)$，不应有显著的差异。否则，很难接受 $X(t)$ 的平稳性假设。

假若 $X(t)$ 是一个正态平稳随机过程，具有方差 σ^2 和相关函数 $\rho(\tau)$。对长为 M 的一段现实序列，由(6.3.1)得到它的 $\bar{x}_i$、s_i^2 和 $r_i(\tau)$。对这些统计量，可分别求出它们的理论方差[4,7]：

$$\left.\begin{aligned}\sigma^2(\bar{x}_i)&=\frac{\sigma^2}{M}\left[1+2\sum_{j=1}^{M}\left(1-\frac{j}{M}\right)\rho(j)\right]\\ \sigma^2(s_i^2)&=\frac{2\sigma^4}{M}\left[1+2\sum_{j=1}^{M}\left(1-\frac{j}{M}\right)\rho^2(j)\right]\\ \sigma^2(r_i(\tau))&=\frac{1}{M-\tau}\left[1+\rho^2(\tau)+2\sum_{j=1}^{M-\tau}\left(1-\frac{j}{M-\tau}\right)(\rho^2(j)+\rho(j+\tau)\rho(j-\tau))\right]\end{aligned}\right\}\tag{6.3.2}$$

用测量序列(6.1.1)给出的 N 个测量数据，由(6.2.5)得到的 σ^2 和 $\rho(\tau)$ 的估计值 s^2、$r(\tau)$ 代入(6.3.2)，给出子序列样本均值 $\bar{x}_i$，方差 s_i^2 和相关函数 $r_i(\tau)$ 的理论方差的渐近估计值 $\bar{\sigma}^2(\bar{x}_i)$、$\bar{\sigma}^2(s_i^2)$ 和 $\bar{\sigma}^2(r_i(\tau))$。取显著水平 $\alpha=0.05$，若

$$\left.\begin{aligned}&|\bar{x}_{i_1}-\bar{x}_{i_2}|>2.77\bar{\sigma}(\bar{x}_i)\\ &|s_{i_1}^2-s_{i_2}^2|>2.77\bar{\sigma}(s_i^2)\qquad\qquad i_1\neq i_2,\quad i_1,\ i_2=1,\ 2,\ \cdots,\ k\\ &|r_{i_1}(\tau)-r_{i_2}(\tau)|>2.77\bar{\sigma}(r_i(\tau))\end{aligned}\right\}\tag{6.3.3}$$

成立，则称差异显著，可以拒绝 $X(t)$ 的平稳性假设。

当 M 取值足够大时，统计量 $\bar{x}_i$、s_i^2、$r_i(\tau)$ 对不同的 i 可近似视为相互独立同分布的随机变量。因此，可以利用连检验(参见本书《蒙特卡洛方法》一章或资料[4])、逆序检验[7]，进一步分析(6.1.1)的平稳性，检验各段子序列的趋势有无异常。

统计检验方法(6.3.3)，利用统计量 $\bar{x}_i$、s_i^2、$r_i(\tau)$ 的估计精度指标，定量地给出了时间序

列平稳性的标准。显然，这一方法比较粗略，当量测序列(6.1.1)的样本量 N 取值不大时，无法检验随机过程 $X(t)$ 的平稳性。

§6.4 非平稳时间序列分析

用§6.3讨论的方法，对时间序列(6.1.1)进行平稳性检验，如果检验的一些统计量差异显著，把 $x(t)$ 作为平稳时间序列进行处理不合理时，利用平稳随机模型(6.2.9)、(6.2.19)、(6.2.27)、(6.2.30)进行分析、预报就很难给出有效的预报结果。因此，有必要根据上述统计分析的结果，对原始数据进行修正，以改进序列的平稳性。

大家熟知，所谓非平稳时间序列，即表示统计性质随时间而异的非常广泛的一类物理数据。在实际问题中遇到的多数物理数据，一般都是非平稳的，只有在简化问题或粗略近似原始问题时，才假定一组原始量测数据是平稳的。因此，对非平稳时间序列进行分析，有着更大的实际意义。

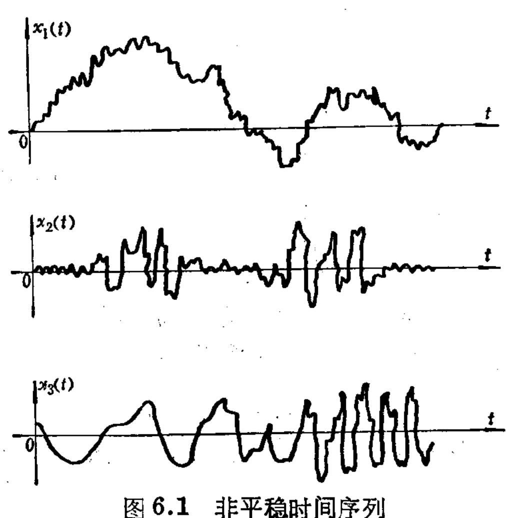


图 6.1 非平稳时间序列

由于非平稳时间序列的一般性和复杂性，至今还没有处理它的一般方法。所以，这里只能讨论其中几类非平稳的时间序列，给出它们的分析和预报方法。

图 6.1 中，给出了几类简单的非平稳时间序列。

从图 6.1 中不难看出，三个时间序列的非平稳特性是不同的。$X_1(t)$ 的量测数据，不在一个固定不变的水平线附近随机摆动，均值 $E[X(t)]$ 将随时间而异。$X_2(t)$ 的量测数据，虽在 $x=0$ 的水平线附近随机摆动，但摆动是不均匀的。可以预见，$X_2(t)$ 的方差 $\sigma^2[X_2(t)]$ 将随时间而异。$X_3(t)$ 的量测数据，有着更为复杂的性质，它的振动频率随时间变化。在数学上，这意味着 $X_3(t)$ 的相关函数将随 t 和 τ 的不同而异。显然，这些非平稳性质的不同组合，可以组成更为复杂的一类非平稳时间序列(如图 6.2)。

设 $D(t)$、$D_1(t)$、$D_2(t)$ 是时间 t 的确定性函数，$\eta(t)$ 是一个各态历经的平稳随机过程。根据上面的分析，通常采用下面几类模型研究非平稳随机过程，作为实际物理过程的近似：

(1) 加法模型 $X_1(t)=D(t)+\eta(t)$；

(2) 乘法模型 $X_2(t)=D(t)\cdot\eta(t)$；

(3) 混合模型 $X_3(t)=D_1(t)+D_2(t)\eta(t)$。

由于乘法模型和混合模型，经过变换，都可化为加法模型进行处理，所以，下面主要讨论可用加法模型

$$X(t)=D(t)+\eta(t)$$

处理的一类随机过程的分析和预报问题。

处理一般非平稳时间序列常用的方法有两种，即参数模型方法和差分模型方法。下面，我们分别讨论这两种方法。

6.4.1 参数模型方法

由 $X(t)$ 的量测数据(6.1.1)，用一些统计分析方法估计函数 $D(t)$ 所含的一些参数[2,4,9]，识别、提取、预报趋势函数项 $D(t)$。我们把这样一类方法，称为参数模型方法。

在时间序列分析中，识别、提取趋势函数项 $D(t)$ 是很重要的一步。在有些问题中，识别、提取 $D(t)$ 本身就是对 $X(t)$ 进行统计分析的最终结果。这里，有一点应特别引起我们的注意：只有在问题的物理机制允许或量测数据本身具有明显的趋势函数项 $D(t)$ 时，进行这种提取才是有意义的。

提取趋势函数项 $D(t)$ 后，把剩余部分

$$\eta(t)=x(t)-D(t) \tag{6.4.1}$$

作为一个平稳时间序列进行分析和预报。综合二者，给出 $X(t)$ 在 $t>t_N$ 时的结果预报值。

为了提取随机过程 $X(t)$ 的趋势函数项 $D(t)$，一般假定 $D(t)$ 有两部分组成，即

$$D(t)=f(t)+p(t) \tag{6.4.2}$$

其中，$f(t)$ 是量测数据(6.1.1)随时间 t 变化的主值函数项，$P(t)$ 是由(6.1.1)可分离出来的周期函数项。

对一般的量测数据(6.1.1)，可以选用相对的时间单位。为了计算处理上的方便，如可以取

$$t_0=1 \quad h=\frac{1}{N} \quad t_n=1+\frac{n}{N} \quad (n=1,\ 2,\ \cdots,\ N)$$

这时，取主值函数项

$$f(t)=c_0+c_1t+c_2t^2+c_3t^3+c_4t^4+c_5t^{-1}+c_6t^{-2}+c_7t^{\frac{1}{2}}+c_8t^{-\frac{1}{2}}+c_9e^{-t}+c_{10}\ln t \tag{6.4.3}$$

拟合量测数据(6.1.1)。用逐步回归算法，给定舍选预报因子 F-检验的临界值 F_1、F_2，让计算机在计算过程中，自动挑选主值函数项 $f(t)$ (6.4.3)的形式和其中的参数，即舍选(6.4.3)中的因子，确定其中的待定参数 c_i。经逐步回归计算，若得到回归系数

$$c_i=0 \quad (i=1,\ 2,\ \cdots,\ 10)$$

则可认为时间序列无主值函数项 $f(t)$。

对一些特殊的量测数据 $x(t)$，根据产生这组数据的物理机制或实际经验，可以选用自己特有的主值函数项 $f(t)$。如对太阳黑子相对数的量测数据进行分析预报时，量测数据(6.1.1)是在一个不到11年的周期内给定的，经分析可用曲线(图 6.2)

$$f(t)=c_0+ct^{\alpha}e^{-\beta t^{\gamma}} \tag{6.4.4}$$

进行拟合。和(6.4.3)相比，(6.4.4)给出了更好的结果。这里，c_0, c, α, β, γ, 是拟合量测数据(6.1.1)时的待定参量，可用《曲线拟合》一章中给出的算法，估计这些参量。

去掉拟合的主值函数项 $f(t)$ 后，得到新的有序数集合

$$v(t)=x(t)-f(t) \quad t=1,\ 2,\ \cdots,\ N$$

对这组数据考虑一个隐含周期模型

$$v(t)=P(t)+\eta(t)$$

这里

$$P(t)=\alpha_0+\sum_{j=1}^{l}(\alpha_j\cos\omega_jt+\beta_j\sin\omega_jt) \tag{6.4.5}$$

其中，a_0、l、α_j、β_j、ω_j 为待定参量。

对随机过程 $X(t)$，根据经验或物理分析，假若已知 $X(t)$ 具有周期 $T_j(j=1, 2, \cdots, k)$，则可象识别、提取主值函数 $f(t)$ 那样，识别、提取人工周期函数项

$$m(t)=a_0+\sum_{j=1}^{k}\left(a_j\cos\frac{2\pi}{T_j}t+b_j\sin\frac{2\pi}{T_j}t\right) \tag{6.4.6}$$

用逐步回归算法，舍选(6.4.6)中的周期 T_j，给出参量 a_j、b_j 的最小二乘估计值，得到人工周期函数项 $m(t)$。

在多数情况下，周期 T_j 是不知道的，需要通过对量测数据 $v(t)$ 的分析，识别、提取由于随机部分 $\eta(t)$ 存在而失真的隐含周期 $T_j=2\pi/\omega_j$。在实际计算中，可以用来进行周期识别的方法是很多的[2,4,9]。下面，我们介绍一种在计算机上应用比较方便的周期图方法。

根据给出的 N 个数据

$$v_1,\ v_2,\ \cdots,\ v_N$$

如果随机过程 $X(t)$ 存在周期，用周期图方法可能分析到的周期有

$$\frac{N}{1},\ \frac{N}{2},\ \cdots,\ \frac{N}{K}$$

这里

$$K=\left[\frac{N}{2}\right]=\begin{cases}\dfrac{N}{2}, & \text{当 } N \text{ 为偶数时}\\[2mm] \dfrac{N-1}{2}, & \text{当 } N \text{ 为奇数时}\end{cases}$$

对可能周期

$$T_k=N/k \quad (k=1,\ 2,\ \cdots,\ K)$$

计算它们的振幅

$$\left.\begin{aligned}\alpha_k&=\frac{2}{N}\sum_{j=1}^{N}v_j\cos\frac{2\pi}{N}kj\\ \beta_k&=\frac{2}{N}\sum_{j=1}^{N}v_j\sin\frac{2\pi}{N}kj\end{aligned}\quad k=1,\ 2,\ \cdots,\ \left[\frac{N-1}{2}\right]\right\} \tag{6.4.7}$$

当 N 为偶数时，有

$$\alpha_{\frac{N}{2}}=\frac{1}{N}\sum_{j=1}^{N}(-1)^j v_j$$

$$\beta_{\frac{N}{2}}=0$$

一般称统计量

$$S_k^2=\frac{1}{2}(\alpha_k^2+\beta_k^2), \quad k=1,\ 2,\ \cdots,\ \left[\frac{N-1}{2}\right]$$

$$S_{\frac{N}{2}}^2=\alpha_{\frac{N}{2}}^2, \qquad \text{当 } N \text{ 为偶数时}$$

为随机过程 $X(t)$ 的周期图。取

$$S^2=\sum_{k=1}^{K}S_k^2$$

可以证明

$$S^2=\frac{1}{N}\sum_{j=1}^{N}(v_j-\bar{v})^2=\sum_{k=1}^{K}S_k^2$$

为了从这些可能周期中选取随机过程 $X(t)$ 的真正周期，给出下面的统计舍选方法。

取
$$S_{i_1}^2=\max\{S_1^2,\ S_2^2,\ \cdots,\ S_K^2\}$$

在 $\eta(t)$ 为高斯白噪声的假定下，统计量

$$y_1=S_{i_1}^2/S^2$$

服从 Fisher 分布

$$P\{y>y_1\}=\sum_{j=0}^{r}(-1)^j C_k^{j+1}[1-(j+1)y_1]^{k-1}$$

其中 r 是使 $1-(r+1)y_1>0$ 成立的最大正整数。

对给定的显著水平 α，若

$$P\{y>y_1\}\geqslant\alpha \tag{6.4.8}$$

可以认为随机过程 $X(t)$ 无周期函数项 $P(t)$。否则，$P\{y>y_1\}<\alpha$，以显著水平 α，接受

$$T_1=\frac{N}{i_1}$$

为随机过程 $X(t)$ 的第一个周期。

取 $S_{i_k}^2$ 为 $S_1^2,\ S_2^2,\ \cdots,\ S_K^2$ 中的第 k 个最大值。统计量

$$y_k=S_{i_k}^2/S^2$$

服从 Fisher 分布

$$F\{y>y_k\}=C_K^{k-1}\sum_{j=0}^{r}(-1)^j C_{K-k+1}^{j+1}\frac{j+1}{j+k}[1-(k+j)y_k]^{K-1}$$

其中 r 是使 $1-(k+r)y_k>0$ 成立的最大正整数。

对给定的显著水平 α，若

$$P\{y>y_k\}<\alpha \tag{6.4.9}$$

接受
$$T_k=N/i_k$$

为随机过程 $X(t)$ 的一个周期。

用蒙特卡洛方法模拟随机模型

$$X(t)=f(t)+P(t)+\eta(t)$$

和分析、预报一些实际问题的经验表明，显著水平 α 不宜取得过大，如一般可取 $\alpha\leqslant 0.01$。

由统计检验(6.4.8)、(6.4.9)，若选得随机过程 $X(t)$ 的 l 个周期 $T_1,\ T_2,\ \cdots,\ T_l$，取圆频率

$$\omega_k=2\pi/T_k=2\pi i_k/N$$

得周期函数项

$$P(t)=a_0+\sum_{k=1}^{l}(\alpha_{i_k}\cos\omega_k t+\beta_{i_k}\sin\omega_k t) \tag{6.4.10}$$

其中

$$a_0=\frac{1}{N}\sum_{j=1}^{N}v_j=\bar{v}$$

利用周期图分析方法，经过(6.4.7)、(6.4.8)、(6.4.9)，识别、提取周期函数项 $P(t)$ (6.4.10)，只能得到以 $\frac{1}{N}$ 为基频的频率分量，无疑，这是不能满足实际问题要求的。引入人工周期函数项 $m(t)$ (6.4.6)，可以弥补周期图分析的这一缺陷。

求得 $f(t)$、$m(t)$、$P(t)$ 后，得到随机过程 $X(t)$ 的趋势函数项

$$D(t)=f(t)+m(t)+P(t)$$

把提取 $D(t)$ 后的剩余

$$\eta(t)=X(t)-D(t) \tag{6.4.11}$$

作为一个平稳随机过程,用§6.2讨论过的方法进行平稳时间序列的分析和预报。

经验表明,选取适当的趋势函数项 $D(t)$,可以改进量测序列(6.1.1)的平稳性,提高预报结果的精度。一般说来,量测数据的样本量 N 比较小,拟合比较复杂的趋势函数项 $D(t)$,容易出现假象,量测数据内部拟合的精度高,预报(外推)的精度低。

根据给出的量测数据(6.1.1),可用§6.3讨论的方法进一步检验由(6.4.11)给出的数据 $\eta(t)$ 的平稳性。假若把 $\eta(t)$ 作为平稳时间序列处理仍不合理时,可作进一步的分析,以改进 $\eta(t)$ 的平稳性。常常根据问题的物理机制和对数据 $\eta(t)$ 进行统计分析的结果,构造一个新的修正算子 H,对 $\eta(t)$ 进行加工,以得到一个渐近的平稳随机过程

$$\varphi(t)=H[\eta(t)]$$

如对太阳黑子相对数进行分析、预报时,提取主值函数项(6.4.4)后,对 $X(t)-f(t)$ 进行统计分析的结果表明,$X(t)-f(t)$ 具有显著的非平稳性(参见图6.2)。因为太阳黑子相对数 $X(t)$,在太阳活动的峰年前后,随机摆动的幅度大,在宁静年份,随机摆动的幅度小,具有 $\eta(t)=X(t)-f(t)$ 和主值函数 $f(t)$ 成正比的特性。根据这一物理特性,选取修正算子

$$H=[f(t)+\alpha]^{-\beta}$$

其中 α、β 为待定参量,且取 $\frac{1}{2}<\beta<1$。经验表明,

$$\varphi(t)=[X(t)-f(t)]/[f(t)+\alpha]^{\beta}$$

较 $\eta(t)=X(t)-f(t)$ 的平稳性有所改进。不难看出,这是用混合模型处理时间序列的一个实例。

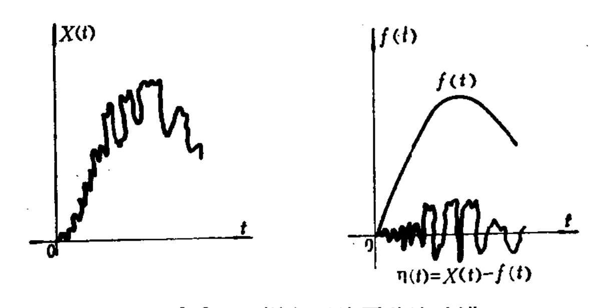


图6.2 时间序列的平稳性改进

在实际计算时,当数据量 N 比较大时,从(6.4.7)可以看出,计算系数 α_i、β_i 的运算量是很大的。这里计算 α_i、β_i 和§6.2中计算平稳过程的谱密度(6.2.7)类似,主要运算量是计算一个有限的傅里叶级数,即计算

$$\sum_{j=1}^{N} z_j \cos j\theta, \quad \sum_{j=1}^{N} z_j \sin j\theta \tag{6.4.12}$$

因此,给出一种比较简单的算法是必要的。

取

$$u_{N+2}=u_{N+1}=0$$

计算递推等式

$$u_j=z_j+2\cos\theta\cdot u_{j+1}-u_{j+2} \quad j=N, N-1, \cdots, 2, 1 \tag{6.4.13}$$

可以证明[1]

$$\sum_{j=1}^{N} z_j \cos j\theta = u_1 \cos\theta - u_2$$

$$\sum_{j=1}^{N} z_j \sin j\theta = u_1 \sin\theta$$

这样，利用(6.4.13)，只需算出三角函数 $\cos\theta$, $\sin\theta$, 便可有效的算出(6.4.12)。

对一些特殊的 N 值，特别在 $N=2^m$(m 为正整数)时，可以利用快速傅里叶算法(FFT)[4,7,10]，更快地得到(6.4.12)。

6.4.2 差分模型方法

这里，非常粗略地介绍一下差分模型方法。

对时间序列$\{x_t\}$，定义：

(1) 后移算子 B

$$Bx_t = x_{t-1}$$

(2) 向后差分算子 ∇、∇_s

$$\nabla x_t = x_t - x_{t-1}$$

$$\nabla_s x_t = x_t - x_{t-s}$$

根据定义，不难证明

$$B^m x_t = B^{m-1} x_{t-1} = \cdots = x_{t-m}$$

$$\nabla x_t = x_t - Bx_t = (1-B)x_t$$

$$\nabla_s x_t = x_t - B^s x_t = (1-B^s)x_t$$

$$\nabla_s = 1 - B^s$$

$$\nabla^{-1} x_t = (1-B)^{-1} x_t = \sum_{j=0}^{\infty} B^j x_t$$

$$= \sum_{j=0}^{\infty} x_{t-j} = \sum_{j=-\infty}^{t} x_j$$

利用后移算子 B 和差分算子 ∇、∇_s，可以表示在§6.2, §6.4.1 中讨论的各种随机模型。

一个 p 阶自回归模型

$$\tilde{x}_t = \phi_1 \tilde{x}_{t-1} + \phi_2 \tilde{x}_{t-2} + \cdots + \phi_p \tilde{x}_{t-p} + \varepsilon_t$$

利用后移算子 B, 可表为

$$\tilde{x}_t = (\phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p)\tilde{x}_t + \varepsilon_t$$

引入 p 阶自回归算子

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

得

$$\phi(B)\tilde{x}_t = \varepsilon_t \tag{6.4.14}$$

如果要求 ε_t 是一个白噪声，$\tilde{x}_t$ 是一个满足自回归模型(6.4.14)的平稳随机过程，则自回归算子 $\phi(B)$ 的系数 ϕ_j 必须满足一定的条件。

将虚拟变量 B 的多项式 $\phi(B)$ 进行分解，得

$$\phi(B) = (1-G_1 B)(1-G_2 B)\cdots(1-G_p B)$$

由(6.4.14),得

$$\tilde{x}_t=\phi^{-1}(B)\varepsilon_t=\sum_{j=1}^{p}\frac{g_j}{1-G_jB}\varepsilon_t=\pi(B)\varepsilon_t \tag{6.4.15}$$

其中

$$\pi(B)=\sum_{j=1}^{p}\frac{g_j}{1-G_jB}$$

称为求和算子。

这时,根据(6.4.15),时间序列 x_t 是以白噪声 ε_t 为输入、传递函数为 $\pi(B)$ 的线性系统的输出函数。因此,只有传递函数 $\pi(B)$ 在虚拟变量 $|B|\leqslant 1$ 时收敛, x_t 才是平稳时间序列。这就要求 p 阶自回归算子 $\phi(B)$ 的特征根,即多项式

$$\phi(B)=0$$

的解都位于单位圆外。根据(6.4.15),这意味着

$$|G_j|<1$$

对 $j=1, 2, \cdots, p$ 同时成立。违反这一条件,就得到各种不同类型的非平稳时间序列。

在实际应用中,出现重根 $|G_j|=1$ 的情形是特别重要的。这时,有简化算子

$$\nabla^d=(1-B)^d$$

$$\nabla_s^D=(1-B^s)^D$$

我们知道,一个平稳随机过程,经过线性差分运算后还是平稳的。一个 $d-1$ 阶多项式

$$f(t)=c_0+c_1t+c_2t^2+\cdots+c_{d-1}t^{d-1}$$

经过 d 阶差分有

$$\nabla^d f(t)=0$$

而一个以整数 S 为周期的周期函数

$$P(t)=P(t+ns),\quad n=\pm1,\ \pm2,\ \cdots$$

经差分

$$\nabla_s P(t)=(1-B^s)P(t)=P(t)-P(t-s)$$

处理,可消去周期分量。

因此,对一个一般的非平稳随机过程

$$X(t)=f(t)+P(t)+\eta(t)$$

其中 $\eta(t)$ 为平稳随机过程,经差分

$$\nabla^d\nabla_s X(t)=\nabla^d\nabla_s\eta(t)$$

处理,得到一个平稳随机过程。

为了大大减少估计参量的个数,提高预报精度,利用 p 阶自回归算子

$$\phi(B)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$$

和 q 阶滑动平均算子

$$\theta(B)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$$

处理平稳过程的分析和预报,得到 $X(t)$ 的一般处理格式:

$$\phi(B)\nabla^d\nabla_s^D X(t)=\theta(B)\varepsilon_t \tag{6.4.16}$$

一般把差分模型(6.4.16)称为 (p, d, q) 阶自回归积分滑动平均模型(ARIMA)。在[5]中,十分详细地讨论了由量测数据(6.1.1)识别、估计、检验随机差分模型(6.4.16)的方法。限于篇幅,这里不再进行更细致的讨论。

§6.5 时间序列分析中的几个问题

在这一节中，很简略地讨论时间序列分析中经常遇到的几个问题，即时间序列的采样、过滤、变换问题，预报结果的精度分析问题。

(一)时间序列采样

给出随机过程 $z(t)$ 的一个现实 $z(t)$。为了在数字计算机上进行分析处理，必须将 $z(t)$ 进行数字离散化。在许多实际问题中，时间序列的采样时间间隔 h 的大小，由问题的物理性质决定。在 h 给定后，得到原始量测数据

$$z_1,\ z_2,\ \cdots,\ z_{N_0} \tag{6.5.1}$$

其中

$$z_n = z(t_0+nh),\quad n=1,\ 2,\ \cdots,\ N_0$$

量测数据(6.5.1)，决定了我们能够分析到的频率范围。由于采样长度 N_0h 和采样时间间隔 h 的限制，分析不到低于 $f_L=1/N_0h$ 和高于 $f_H=1/2h$ 的频率分量，形成采样误差。

(二)时间序列滤波

在时间序列分析中，把比较广泛的一类数据处理方法叫做数字滤波。数字滤波可以用来平滑原始数据、降低量测误差、减少处理的数据量，实现数据分频以研究分频数据的性质。

给出原始量测数据(6.5.1)，常用的简单滤波方法有：

(1) 等权平滑滤波

$$y_n=\frac{1}{n_1}\sum_{j=1}^{n_1} z_{r+n_2(n-1)+j},\quad n=1,\ 2,\ \cdots,\ N$$

这里，参量 $n_1\geqslant n_2\geqslant 1$

$$N=\left[\frac{N_0-n_1+n_2}{n_2}\right],\quad r=N_0-n_1+n_2-Nn_2$$

(2) 加权滤波

$$y_n=\sum_{j=1}^{n_3} h_j z_{n-1+j},\quad n=1,\ 2,\ \cdots,\ N$$

这里，h_j 为给定的滤波权函数，$n_3>1$，$N=N_0-n_3+1$

(3) 累积滤波

$$y_0=0,\quad y_n=y_{n-1}+z_n=\sum_{j=1}^{n} z_j$$

$$n=1,\ 2,\ \cdots,\ N\quad N=N_0$$

(4) 低通递推滤波

$$y_0=0,\ y_n=\alpha y_{n-1}+(1-\alpha)z_n$$

$$n=1,\ 2,\ \cdots,\ N,\quad N=N_0,\quad 0<\alpha<1$$

(三)时间序列变换

为了扩大时间序列处理物理数据的范围，对经过简单滤波处理后的数据 y_n 引进一类非线性变换，用含有一个或多个参量的数据 x_n 代替 y_n 进行分析。变换的类型可以根据问题的物理性质和数据自身的特性进行选取。如可取

$$x_n=\begin{cases}[ay_n+b]^{\lambda}, & 当\ \lambda\neq 0\ 时\\ \ln[ay_n+b], & 当\ \lambda=0\ 时\end{cases}$$

其中变换参量 a、b, 在 $\lambda\neq 1$ 时,应使

$$ay_n+b>0$$

对一切 n 成立。在参量 $a=\lambda=1$, $b=0$ 时, $x_n=y_n$, 量测数据保持不变。

(四)预报精度分析

对时间序列(6.1.1)进行统计分析, 用回归模型(6.4.3)识别、提取 $X(t)$的主值函数项 $f(t)$,用调和分析模型(6.4.5)、(6.4.6)识别、提取 $X(t)$的周期函数项 $P(t)$、$m(t)$, 用自回归模型(6.2.9), (6.2.27)给出平稳随机过程 $\eta(t)$的预报值 $\eta^*(N+l)$,最后综合给出预报结果

$$x^*(N+l)=f(N+l)+m(N+l)+P(N+l)+\eta^*(N+l)$$

形式上完成了 $X(t)$的分析、预报工作。但是,对实际应用来讲,仅仅给出 $x(N+l)$的预报值 $x^*(N+l)$还是远远不够的。只有在一定信度条件下给出 $x^*(N+l)$的预报精度,才能知道预报值 $x^*(N+l)$能否在实际中应用。

在时间序列分析中,一般采用后验预报方法进行预报结果的精度分析。这时,把已经采用的预报方案施用于已有的量测数据上,用远期的已知数据预报近期已有的结果,给出后验预报残差 δ。

如用 $N-k$ 个量测数据

$$x_1,\ x_2,\ \cdots,\ x_{N-k}$$

进行分析、预报,得到 x_{N-k+1} 的预报值 x^*_{N-k+1}, 求得预报残差

$$\delta_k=x_{N-k+1}-x^*_{N-k+1}$$

对 $k=1,\ 2,\ \cdots,\ K$, 得到一组后验预报残差

$$\delta_1,\ \delta_2,\ \cdots,\ \delta_K \tag{6.5.2}$$

对数据组(6.5.2)进行分析,给出它们的均值、方差和经验分布函数,得到预报的精度指标。

如果进行的后验预报具有代表性,在物理条件不变的情况下,后验预报残差 δ 的分布和今后实际预报误差的分布,在统计上应该是一致的。因此,根据后验预报的精度可以推断实际预报的可靠性。

下面,给出预报精度分析的几个统计指标。

对量测数据(6.1.1),计算原始方差

$$S_1^2=\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})^2$$

由后验预报残差序列(6.5.2),计算统计量

$$\bar{\delta}=\frac{1}{K}\sum_{k=1}^{K}\delta_k$$

$$S_2^2=\frac{1}{K}\sum_{k=1}^{K}(\delta_k-\bar{\delta})^2$$

给出后验标准差的比值

$$C=s_2/s_1$$

和或然误差的观测频率

$$P\{|\delta_k-\bar{\delta}|<0.6745S_1\}$$

据此,可以给出预报方案好坏的一个参考标准。

预 报 精 度	P	C
好	>0.95	<0.35
合 格	>0.80	<0.50
勉 强	>0.70	<0.65
不合格	⩽0.70	⩾0.65

根据上面讨论的结果，用图6.3所示的框图给出时间序列分析的一般过程。

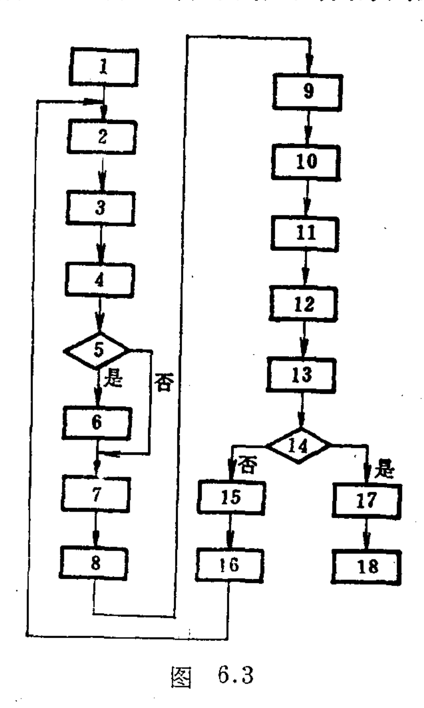


图 6.3

框 图 说 明

[框1] 输入原始量测数据及控制参量

(1) 原始量测数据

$$z(t) \quad t=1, 2, \cdots, N_0$$

(2) 基本模型

(3) 控制参量

[框2] 原始数据线性滤波

$$y(t)=L_1[z(t)] \quad t=1, 2, \cdots, N<N_0$$

[框3] 滤波结果进行变换

$$x(t)=L_2[y(t)] \quad t=1, 2, \cdots, N$$

[框4] $x(t)$的主值函数项$f(t)$的识别、提取、预报

[框5] 是否给出人工周期T_j

[框6] 根据周期T_j，进行人工周期函数项的识别、提取、预报

[框7] 进行自然周期函数项$P(t)$的识别、提取、预报

[框8] 计算

$$D(t)=f(t)+m(t)+P(t) \quad t=1, 2, \cdots, N, N+1, \cdots, N+L$$

$$\eta(t)=x(t)-D(t) \quad t=1, 2, \cdots, N$$

[框 9] 进行 $\eta(t)$ 的平稳性改进

$$\varphi(t)=L_3[\eta(t)] \quad t=1, 2, \cdots, N$$

[框 10] 对 $\varphi(t)$ 进行平稳性分析

[框 11] 对 $\varphi(t)$ 构造随机模型，进行 $\varphi(t)$ 的统计预报，给出预报值 $\varphi^*(N+l)(l=1, 2, \cdots, L)$

[框 12] 给出综合预报值

$$\left.\begin{aligned} x^*(N+l)&=D(N+l)+L_3^{-1}[\varphi^*(N+l)] \\ y^*(N+l)&=L_2^{-1}[x^*(N+l)] \end{aligned}\right\} \quad l=1, 2, \cdots, L$$

[框 13] 进行结果预报的精度分析

[框 14] 预报精度是否满足要求？

[框 15] 预报精度不满足要求，停机后决定是否改变模型和控制参数

[框 16] 根据上面分析的结果，给出新的模型和参量，转框 2 进行新的计算

[框 17] 输出分析结果和预报结果

[框 18] 停机

参 考 资 料

[1] A. 拉尔斯登, H. S. 维尔夫著, 徐献瑜等译, «数字计算机上用的数学方法» 17. 多重回归分析, p. 302-316, 19. 自相关及谱分析, p. 331-342, 24. Fourier 分析, p. 402-408, 科学技术出版社, 1963。

[2] U. 格列南特, M. 罗逊勃列特著, 郑绍濂等译, «平稳时间序列的统计分析», 科学技术出版社, 1962。

[3] M. 费史著, 王福保译, «概率论及数理统计», 科学技术出版社, 1962。

[4] Bendat J. S., Piersol A. G., "Random Data: Analysis and Measurement Procedures", John Wiley & Sons, 1971.

[5] Box G. E. P., Jenkins G. M., "Time Series Analysis Forecasting and Control", Holden-Day, 1970.

[6] Jones R. H., "Prediction of multivariate time series", J. appl. meteer., 3, 1964, 285-289.

[7] Otnes R. K., Enochson L., "Digital Time Series Analysis", John Wiley & Sons, 1972.

[8] French W. F., "Weighting coefficients for the prediction of stationary time series from the finite past", SIAM. J. Appl. Math., 15(6), 1967, 1502-1510.

[9] Wilks W. W., "Mathematical Statistics", 17. time series, 514-539, John Wiley & Sons, 1962.

[10] Чеголин П. М., Пойда В. Н., "Методы, Алгоритмы и Програмны Статистического Анализа", «Наука и Механика», 1971.

第七章　蒙特卡洛方法

§7.1　概　　论

蒙特卡洛(Monte Carlo)方法，是一类通过随机变量的统计试验、随机模拟，求解数学物理、工程技术问题近似解的数值方法。在资料[1、11、13、14、15]中，也把这类方法，叫做统计试验方法，随机模拟方法。

下面，我们通过一维积分

$$\theta=\int_0^1 f(x)dx \quad (0<f(x)<1) \tag{7.1.1}$$

的随机模拟，说明用蒙特卡洛方法求解一般实际问题的基本步骤和主要特点。

为了用蒙特卡洛方法模拟积分(7.1.1)，就需要根据积分(7.1.1)的特点，构造一个今后简称为概型的数学概率模型。

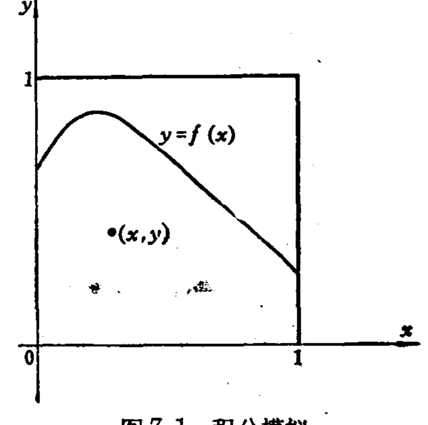


图7.1　积分模拟

在图7.1中，正方形内曲线 $y=f(x)$ 下面的面积等于积分值 θ。如果在正方形内任投一点 (x, y)，则随机点 (x, y) 位于曲线 $f(x)$ 下面的概率

$$P=P\{y<f(x)\}=\int_0^1\int_0^{f(x)} dy\,dx$$

等于积分值 θ。据此，可以构造模拟积分(7.1.1)的概型。

按(0, 1)上的均匀分布律，产生随机变量 (x, y) 的抽样值(参见§7.2)，作为正方形内随机点的坐标，模拟随机投点试验。取随机变量 η。当随机点 (x, y) 位于曲线 $f(x)$ 的下面时，随机投点试验成功，η 取值为1，否则，取值为0，即

$$\eta=\begin{cases}1, & y<f(x),\ \text{随机投点试验成功}\\ 0, & y>f(x),\ \text{随机投点试验失败}\end{cases}$$

大量产生相互独立随机变量 (x, y) 的抽样值 (x_i, y_i)，得到 η 的观测值 $\eta_i=\eta(x_i, y_i)$。在 N 次随机投点试验中，$\sum_{i=1}^{N}\eta_i$ 给出随机投点试验成功的总次数。

根据统计估计理论和中心极限定理[2]，样本均值

$$\bar{\eta}=\frac{1}{N}\sum_{i=1}^{N}\eta_i \tag{7.1.2}$$

给出积分值 θ 的无偏估计，样本标准差

$$S_\eta=\left[\frac{1}{N}\sum_{i=1}^{N}(\eta_i-\bar{\eta})^2\right]^{1/2} \tag{7.1.3}$$

给出 $\bar{\eta}$ 精度的统计估计。

实际问题中的概型总是复杂的。一般，象用(7.1.2)对积分(7.1.1)进行直接模拟，结果

的精度很低，往往是不成功的。对多数概型，必须根据它们的特点，设计和使用降低模拟方差、加速结果收敛的方法。如模拟积分(7.1.1)时，产生(0, 1)上均匀分布的随机变量 x_i，可以得到比(7.1.2)更为有效的估计(参见§7.7)

$$\theta_1=\frac{1}{N}\sum_{i=1}^{N}f(x_i) \tag{7.1.4}$$

从积分(7.1.1)的模拟过程中可以看出，用蒙特卡洛方法模拟一个实际问题，基本步骤有：

(1) 根据实际问题，构造模拟概型；

(2) 根据概型的特点，设计、使用降低方差的各类方法，加速模拟结果的收敛；

(3) 给出概型中各种不同分布随机变量的抽样方法；

(4) 统计处理模拟结果，给出问题的解和精度估计。

这里给出的模拟步骤，组成了蒙特卡洛方法研究的基本课题，即蒙特卡洛方法的基本理论，随机变量的产生、检验及其实际应用。

把随机模拟方法用于近似数值计算领域已有近百年的历史。由于模拟试验工具的限制，除很个别的例子(如通过蒲丰投针概型计算圆周率 π[2]，验证 t-分布等)，很少有人用来求解实际问题。

随着科学技术的发展，在生产斗争、科学实验过程中，出现了许多复杂难解的问题，如核物理中描述质点运动的迁运方程[1,14]，大型系统的可靠性分析[10]，地震波的模拟试验[6]，高维数学物理问题求解，多元统计分析[1,7,14]，医学、技术中的诊断、识别，大型系统模拟试验，大规模生产过程中的运筹规划[13,14]等等。对这些问题用传统的物理试验或数学方法进行处理，常常感到十分困难，而用蒙特卡洛模拟，常使人眼界开扩，找到新的处理方法。快速数字计算机的出现和发展，为蒙特卡洛方法提供了强有力的模拟工具，使这一方法得到愈来愈广泛的应用[4,8,15]。

从积分(7.1.1)的模拟过程可知，蒙特卡洛方法是模拟概型中的一个随机变量 η，更确切地说，是模拟一个复杂的随机变量 $x_1, x_2, \cdots, x_m$ 的函数

$$\eta=\eta(x_1, x_2, \cdots, x_m)$$

通过 η 的随机模拟，得到抽样值 $\eta_1, \eta_2, \cdots, \eta_N$，统计处理后，给出 η 的概率分布或各阶矩的估计值，得到概型的解。在实际问题中，经常要求给出 η 的数学期望 $E(\eta)$ 和标准差 σ_η 的估计值，即(7.1.2)、(7.1.3)中的 $\bar{\eta}$ 和 S_η。因此，模拟概型的随机性，模拟算法的简单性，是蒙特卡洛方法的第一个特点。

(7.1.2)、(7.1.4)两种不同的模拟算法，都能给出积分(7.1.1)的无偏估值。模拟过程的这种灵活性，组成蒙特卡洛方法的另一特点。

蒙特卡洛方法的误差估计和收敛性，别具风格，构成它的第三个特点。

设模拟的随机变量 $\eta(x_1, x_2, \cdots, x_m)$ 具有有限的方差 σ^2，它的样本均值

$$\bar{\eta}=\frac{1}{N}\sum_{i=1}^{N}\eta_i$$

给出概型的解。当样本量 N 充分大时，模拟结果的误差

$$|\bar{\eta}-E(\eta)|<\varepsilon \tag{7.1.5}$$

以概率

$$P=\int_{\frac{-\varepsilon\sqrt{N}}{\sigma}}^{\frac{\varepsilon\sqrt{N}}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{7.1.6}$$

成立。取置信水平 $P=95.5\%$，误差

$$\varepsilon=2\sigma/\sqrt{N} \tag{7.1.7}$$

由(7.1.5)、(7.1.6)、(7.1.7)可以看出，蒙特卡洛模拟的精度和收敛过程，具有如下特点：

(1) 收敛过程服从概率规律。结果精度 $\sigma/\sqrt{N}$ 象物理实验一样，可在模拟过程中进行估计；

(2) 模拟结果的精度和概型的维数 m 无关，因而蒙特卡洛方法特别适宜求解高维问题；

(3) 收敛速度 $O(1/\sqrt{N})$；

(4) 降低方差 σ^2，地位重要，效果显著。

蒙特卡洛方法以 $O(1/\sqrt{N})$ 的速度收敛，理论上已无法改善，和一般数值方法相比是比较慢的。这样的收敛速度，要把结果的精度提高一位，就要百倍地增加模拟工作量。在实际应用中，可精心设计模拟概型，改进抽样方法，降低方差 σ^2。理论分析和实际应用的经验表明，综合利用各类降低方差的技巧，可十倍，甚至百倍地提高模拟结果的收敛速度。

§7.2 随机数的产生

用蒙特卡洛方法模拟一个实际问题时，用到各种不同分布的随机变量。在理论上，只要有了一种连续分布的随机变量，通过变换、舍选等抽样方法，就可以得到任意分布的随机变量(参见§7.3、§7.4、§7.5)。在连续分布函数中，(0, 1)上的均匀分布函数

$$F(x)=\begin{cases}0, & x\leqslant 0\\ x, & 0<x\leqslant 1\\ 1, & 1<x\end{cases}$$

是最简单的。因此，在蒙特卡洛模拟中，多是先产生均匀分布随机变量 R 的抽样值 $r_i(i=1, 2, \cdots)$，通过变换、舍选各类方法，再产生其它分布随机变量的抽样值。为简单计，我们把(0, 1)上均匀分布随机变量 R 的抽样值 r_i 叫做随机数。

在电子计算机上，已经使用的产生随机数的方法，可大致分为三类：

(1) 把已有的随机数表(如资料[12])输入机器；

(2) 用物理方法(如噪声型随机数发生器[13,14])产生真正的随机数；

(3) 用数学方法产生伪随机数。

产生随机数的数学方法是利用数字计算机的运算性能，根据递推公式

$$r_{n+k+1}=g(r_{n+1}, r_{n+2}, \cdots, r_{n+k}) \tag{7.2.1}$$

由程序直接产生数值序列 $\{r_n\}$。这种产生随机数的方法速度快，占用机器的内存小，对模拟的问题可以进行复算检查，故发展快、使用广[5,8]。

由于数字计算机上表示一个数字的字长有限，故只能表示有限个不同的数。所以，严格说来，在计算机上不能产生真正连续分布的随机数。象一般数值计算中用差分代替微分那样，在蒙特卡洛模拟中，用离散分布的随机数代替连续分布的随机数。因此，用递推方法(7.2.1)产生的数值序列 $\{r_n\}$ 是完全确定的，到一定长度周而复始，出现周期现象。这些、和

随机数的基本性质都是矛盾的。所以，在机器上，用数学方法，不可能产生真正的随机数。但是，只要产生的数值序列$\{r_n\}$能够通过随机数的各类统计检验(参见§7.6)，就可以把它们当作真正的随机数使用。为和真正的随机数相区别，通常把用数学方法产生的随机数叫做伪随机数。

在产生伪随机数的数学方法中，有迭代取中法，移位法和同余法[14]。下面，我们介绍统计性质较好，使用较广的两类同余法：乘同余法和混合同余法。

用混合同余法产生伪随机数的递推同余式是

$$x_{n+1}\equiv\lambda x_n+C \pmod M \tag{7.2.2}$$

其中，初值x_0，乘子λ，增量C和模M取非负整数。当$C=0$时，有

$$x_{n+1}\equiv\lambda x_n \pmod M \tag{7.2.3}$$

是用乘同余法产生伪随机数的递推同余式。这里

$$A\equiv B \pmod M$$

表示以M为模的同余式，$A<M$，取B被M整除后的余数。取

$$r_n=x_n/M \tag{7.2.4}$$

作为产生的随机数。显然，$0\leqslant r_n<1$。在计算机上，

$$x_{n+1}\equiv\lambda x_n \pmod M$$

可用双倍位乘法运算，取后面的尾数部分来实现。

下面，从产生伪随机数的基本要求——周期长、产生的速度快、统计性质优(如随机数的均匀性、独立性、随机性等，§7.6)——讨论参数x_0，λ，C，M的选取问题[5]。

在二进制数字计算机上，取$M=2^l$，其中l是数字计算机上一个数字尾部的字长。这时，用(7.2.2)产生随机数，最大可能周期$T=2^l$，而(7.2.3)的最大可能周期$T=2^{l-2}$。

为了使序列x_n取到这样大的周期，应选取参数：

(1) 在混合同余法(7.2.2)中，取

$$\lambda=4q_1+1,\quad C=2a_1+1,\quad x_0\text{为任意非负整数}$$

(2) 在乘同余法(7.2.3)中，取

$$\lambda=8q_2\pm3,\quad x_0=2a_2+1$$

这里，q_1、q_2、a_1、a_2，均取正整数。

理论分析和统计检验表明，λ取值过小或它的二进制形式中0、1呈规则性排列时，都不能产生统计性质理想的数值序列，反之，一般是可取的。特别，对乘同余法(7.2.3)，取

$$\lambda=5^{2s+1}$$

是成功的。其中参量s是使$5^{2s+1}<2^l$成立的最大正整数。如在$M=2^{31}$时，可取$\lambda=5^{13}$。

在用蒙特卡洛方法模拟复杂概型时，使用乘同余法产生的伪随机数序列，还存在一些缺点。实际模拟提出对同余法改进的要求，组合同余法便是其中的一个。

用乘同余法(7.2.3)产生m个随机数

$$r_1,\ r_2,\ \cdots,\ r_m \tag{7.2.5}$$

用混合同余法(7.2.2)产生另一随机数，从(7.2.5)中随机地选取一个，作为实际中使用的随机数。这样，从乘同余法产生的随机序列$\{r_n\}$中，随机地选取其中一个子列$\{r_{n_k}\}$，它的统计性质，经统计检验优于原序列。

在机器上进行模拟计算时，先产生m个乘子

$$\lambda_1, \lambda_2, \cdots, \lambda_m$$

其中

$$\lambda_j \equiv \lambda^j \pmod M$$
$$j=1, 2, \cdots, m$$

用混合同余法产生的随机数来随机选取 λ_j。取

$$x_{n_{k+1}} \equiv \lambda_j x_{n_k} \pmod M$$

把 $\left\{\frac{x_{n_k}}{M}\right\}$ 取为模拟用的随机数列。这样，在实际计算中，并不需要把 $r_1, r_2, \cdots, r_m$ 如实地产生出来，以提高机器的产生效率。在二进制数字计算机上，可取 $m=2^5$ 或 2^6。

§7.3 随机变量抽样

在得到(0, 1)上均匀分布随机数序列 $\{r_n\}$ 之后，必须给出概型中各种不同分布随机变量的抽样方法，才能进行蒙特卡洛模拟。下面，分别讨论离散随机变量、连续随机变量的抽样方法。

7.3.1 离散随机变量抽样

设随机变量 η 以概率 $P_1, P_2, \cdots$, 分别取值 $a_1, a_2, \cdots$, 即

$$P\{\eta=a_n\}=P_n, \quad n=1, 2, \cdots$$

称为离散的。这里，$0<P_n<1$, $\sum_n P_n=1$。取

$$P^{(0)}=0, \quad P^{(n)}=\sum_{i=1}^{n} P_i, \quad n=1, 2, \cdots$$

由随机数序列 $\{r_i\}$ 中，依次选取随机数 r_i, 求满足检验条件

$$P^{(n-1)} \leqslant r_i < P^{(n)} \tag{7.3.1}$$

的 n 值。这时，得到随机变量 η 的抽样值

$$\eta=a_n$$

对一些特殊的离散分布，如几何分布、二项分布、超几何分布、泊松分布等，从它们的概率意义出发，可以得到和(7.3.1)不同的一些抽样方法[11]。

例1 产生服从泊松分布的随机变量 η。

随机变量 η 服从离散分布

$$P\{\eta=n\}=e^{-\lambda}\frac{\lambda^n}{n!}, \quad n=0, 1, 2, \cdots; \quad \lambda>0 \tag{7.3.2}$$

称为泊松分布的。对随机数序列 $\{r_i\}$，求满足条件

$$\prod_{i=0}^{n} r_i \leqslant e^{-\lambda} < \prod_{i=0}^{n+1} r_i$$

的 n 值。易证，$\eta=n$, 服从泊松分布(7.3.2)。

7.3.2 连续随机变量抽样

下面，讨论几类具有分布函数 $F(x)$，密度函数 $f(x)$ 的连续随机变量的抽样方法。

(一)直接抽样

若随机变量 η 具有连续的分布函数 $F(x)$，则随机变量

$$R=F(\eta) \tag{7.3.3}$$

均匀分布在(0, 1)上。(7.3.3)给出了两个随机变量之间的基本关系，是由随机数 r 对 η 进行直接抽样的依据。它的直观意义如图7.2所示。

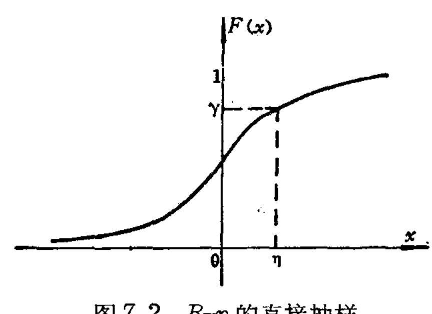


图7.2 R-η 的直接抽样

例2 产生指数分布

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

的随机变量 η。

由(7.3.3)

$$R=\int_{-\infty}^{\eta}f(x)dx=1-e^{-\lambda\eta}$$

得

$$\eta=-\frac{1}{\lambda}\ln(1-R)$$

因 R 和 $(1-R)$ 同分布，故有

$$\eta=-\frac{1}{\lambda}\ln R$$

它在蒙特卡洛模拟中，使用较广，经常用来描述电子元件的稳定时间，系统的可靠性和质点的游动自由程。

(二)变换抽样

在蒙特卡洛方法中，研究随机变量的函数变换，特别是二维均匀分布随机变量的函数变换，为产生一般随机变量提供速度更快的方法。

例3 用变换抽样，产生标准正态分布的随机变量 u。

若随机变量 η 以

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{7.3.4}$$

为密度函数，称为正态的，简记为 $\eta\sim N(\mu,\sigma^2)$。特别，当 $E(\eta)=\mu=0, D(\eta)=\sigma^2=1$ 时，称为标准正态的，并用 u 表示，即 $u\sim N(0, 1)$。

正态随机变量 η，在统计分析和蒙特卡洛模拟中，有着广泛的应用。u 和 η 之间存在关系

$$\eta=\sigma u+\mu$$

取随机数 r_1, r_2，利用二元函数变换[9]

$$\begin{aligned}u_1&=\sqrt{-2\ln r_1}\cos 2\pi r_2\\ u_2&=\sqrt{-2\ln r_2}\sin 2\pi r_2\end{aligned} \tag{7.3.5}$$

得到两个相互独立的 $N(0, 1)$ 分布随机变量 u 的抽样值。

解方程(7.3.5)，得

$$\begin{cases}r_1=\exp\left\{-\frac{1}{2}(u_1^2+u_2^2)\right\}\\ r_2=\frac{1}{2\pi}\operatorname{arc\,tg}\frac{u_1}{u_2}\end{cases}$$

故知随机变量 u_1、u_2 的密度函数

$$f(x,y)=g(r_1,r_2)|J|=\frac{1}{2\pi}\exp\left\{-\frac{1}{2}(x^2+y^2)\right\}=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}$$

其中，J 为变换 r_1、r_2 的 Jacobi 行列式。故知，u_1、u_2 相互独立，服从 $N(0, 1)$分布。

（三）舍选抽样

产生随机变量的舍选方法，利用满足一定的检验条件进行补偿，其方法灵活、计算简单、使用较广。下面介绍几种不同的舍选抽样方法，最后给出它们的一般性证明。

舍选方法 I. 设 η 是(0, 1)上以

$$f(x)=Lh(x),\quad (0<x<1)$$

为密度函数的随机变量。其中，$L=\sup f(x)$，取有限值，$0\leqslant h(x)\leqslant 1$。抽样算法可用框图 7.3 表示：

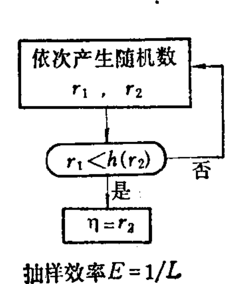


图 7.3　抽样框图

图 7.4　舍选抽样

舍选抽样方法 I 的直观意义，可用图 7.4 说明。

在边长为 1 和 L 的矩形内任投一点 P。若随机点 P 位于曲线 $y=f(x)$ 的下面时(如 P_1)，即满足条件

$$Lr_1<f(r_2)$$

取它的横坐标 r_2 为模拟的随机变量 η，否则，拒绝该点(如 P_2)，进行新的试验。

在一次试验中，随机点 P 位于曲线 $y=f(x)$下面的概率

$$E=P\{Lr_1<f(r_2)\}=\int_0^1 dx\int_0^{f(x)/L} dy=\frac{1}{L}$$

称为舍选方法的抽样效率。抽样效率的倒数，$L=\frac{1}{E}$，是产生一个随机变量 η 的平均试验次数，可用来比较不同舍选抽样方法的好坏。

例 4　产生具有密度函数

$$f(x)=\begin{cases}2x, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

的随机变量 η。

用直接抽样(7.3.3)，得随机平方根

$$\eta=\sqrt{r}$$

进行抽样时，用到开方子程序，计算量比较大。利用舍选方法 I，考虑到随机数 r_1、r_2 的对称性，则有

$$\eta=\max\{r_1,\ r_2\}$$

舍选方法 II. 设随机变量 η 的密度函数 $f(x)$ 可表为

$$f(x)=Lh(x)g(x)\qquad(7.3.6)$$

其中，常量 $L>1$，$0\leqslant h(x)\leqslant 1$，$g(x)$是随机变量$\xi$的密度函数。随机变量$\eta$的抽样算法可用框图7.5表示：

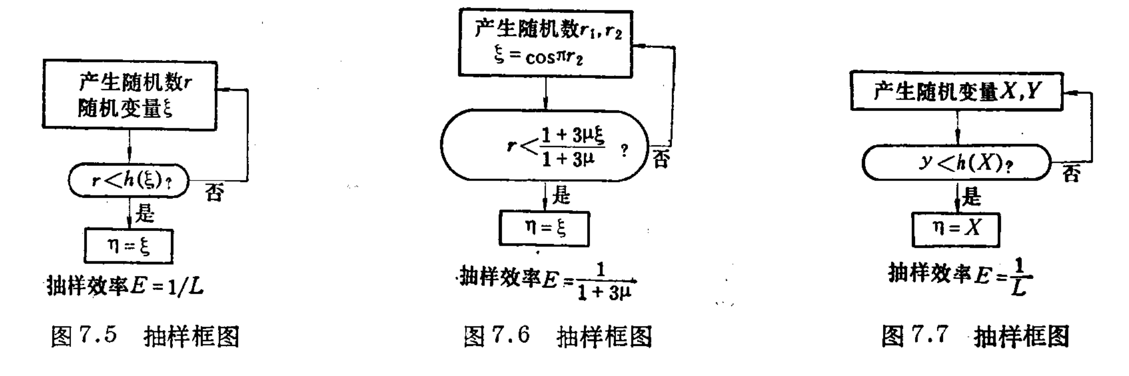


图7.5　抽样框图　　图7.6　抽样框图　　图7.7　抽样框图

例5　产生具有密度函数

$$f(x)=\begin{cases}(1+3\mu x)\dfrac{1}{\pi\sqrt{1-x^2}}, & |x|<1\text{ 时}\\ 0, & \text{其它}\end{cases}$$

的随机变量η。这里，参量$\mu>0$。

根据舍选抽样方法II，按照(7.3.6)分解$f(x)$。这里，

$$L=1+3\mu,\qquad h(x)=\frac{1+3\mu x}{1+3\mu}$$

$$g(x)=\begin{cases}\dfrac{1}{\pi\sqrt{1-x^2}}, & |x|<1\text{ 时}\\ 0, & \text{其它}\end{cases}$$

是随机变量ξ的密度函数。用直接抽样，得随机余弦$\xi=\cos\pi r$(例6将给出更快的抽样方法)。由此，得η的抽样算法如图7.6所示。

舍选方法III. 设随机变量η的密度函数

$$f(x)=L\int_{-\infty}^{h(x)}g(x,\ y)dy$$

其中，$g(x,y)$是二维随机向量(X,Y)的联合密度函数，函数$h(x)$在y的定义域上取值，常量

$$L=\frac{1}{\int_{-\infty}^{+\infty}\int_{-\infty}^{h(x)}g(x,\ y)dy\,dx}>1$$

是规格化常数。η有如框图7.7所示的抽样算法。

当随机变量X、Y相互独立时，有

$$f(x)=Lg_1(x)\int_{-\infty}^{h(x)}g_2(y)dy$$

显然，舍选抽样方法I、II，是III的特例。对舍选方法III证明如下：

$$P\{\eta<x\}=P\{X<x|Y<h(X)\}=\frac{P\{X<x,\ Y<h(X)\}}{P\{Y<h(X)\}}$$

$$=\frac{\int_{-\infty}^{x}dt_1\int_{-\infty}^{h(t_1)}g(t_1,\ t_2)dt_2}{\int_{-\infty}^{+\infty}dt_1\int_{-\infty}^{h(t_1)}g(t_1,\ t_2)dt_2}=\int_{-\infty}^{x}\left[L\int_{-\infty}^{h(t_1)}g(t_1,\ t_2)dt_2\right]dt_1$$

例 6　各向同性散射方位角 φ 的余弦抽样。

质点在各性同性散射过程中，方位角 φ 均匀分布在 $(0, 2\pi)$ 上。直接抽样方向余弦，有

$$\begin{cases} \eta = \cos 2\pi r \\ \xi = \sin 2\pi r \end{cases}$$

用到两个标准子程序，运算量大。不难计算，随机变量 η 有密度函数

$$f(x) = \begin{cases} \dfrac{1}{\pi} \dfrac{1}{\sqrt{1-x^2}}, & |x|<1 \text{ 时} \\ 0, & \text{其它} \end{cases}$$

定义随机变量

$$\begin{cases} X = \dfrac{r_1^2 - r_2^2}{r_1^2 + r_2^2} \\ Y = r_1^2 + r_2^2 \end{cases}$$

有联合密度函数

$$g(x, y) = \begin{cases} \dfrac{1}{4\sqrt{1-x^2}}, & |x|<1,\ 0<y<\dfrac{2}{1+|x|} \\ 0, & \text{其它} \end{cases}$$

这时

$$f(x) = \frac{4}{\pi}\int_{-\infty}^{1} g(x, y)\,dy$$

根据舍选抽样 III，有如图 7.8 所示的抽样算法。

(四)近似抽样

上面讨论的几种抽样方法，从理论上讲是精确的，即除去用伪随机数代替随机数形成的误差外，不含系统误差。这里讨论的近似抽样算法(分布近似、密度近似或概率近似)，由于所选方法的性质而含有系统误差。在实际模拟中，一般说来都不影响结果的精度。

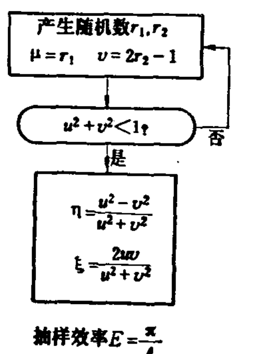


图 7.8　抽样框图

设随机变量 η 以连续函数 $F(x)$ 为分布函数。根据 (7.3.3)，$R = F(\eta)$。当 $F(x)$ 的逆函数

$$\eta = G(R)$$

不易求出时，用直接抽样是困难的。但我们知道，在多数情况下，$G(R)$ 具有如下性质：

当 $R \to 0$ 时，$G(R) \to -\infty$；当 $R \to 1$ 时，$G(R) \to +\infty$。因此，可以利用最小二乘法拟合曲线 $G(R)$。取拟合函数

$$G(R) = a + bR + cR^2 + \alpha(1-R)^2 \ln R + \beta R^2 \ln(1-R) \qquad (7.3.7)$$

对非常广泛的一类分布函数是可行的。其中系数 a、b、c、α、β 为待定参数。当然，对 $G(R)$ 也可采用其它的逼近方法，如有理分式逼近[9]。

例 7　拟合曲线 $\eta(R)$，满足方程

$$R = \int_{-\infty}^{\eta(R)} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

即用近似抽样，产生 $N(0, 1)$ 的正态随机变量。

取点　　$R_k = \dfrac{k}{200} \quad (k = 1, 2, \cdots, 199)$

根据(7.3.7)，用逐步回归算法，得系数

$$a=-0.8368,\quad b=1.6736=-2a,\quad c=0$$

$$\alpha=0.3315,\quad \beta=-0.3315=-\alpha$$

例 8　用概率近似，产生 $N(0,\ 1)$分布的随机变量 u。

根据中心极限定理，取随机数 $r_1,\ r_2,\ \cdots,\ r_n$，有概率渐近抽样：

$$u=\sqrt{12n}\left(\frac{1}{n}\sum_{i=1}^{n}r_i-\frac{1}{2}\right)$$

在实际应用中，多取 $n=6$ 或 $n=12$。

在 $n=6$ 时，
$$u=\sqrt{2}\sum_{i=1}^{3}(r_{2i}-r_{2i-1})$$

在 $n=12$ 时，
$$u=\sum_{i=1}^{6}(r_{2i}-r_{2i-1})$$

(五)复合抽样

产生随机变量的复合抽样方法，是把要进行抽样的密度函数 $f(x)$，分解为一些经过适当选取的密度函数 $f_n(x)$的概率和。在数学上，复合抽样可表为：

$$f(x)=\sum_n P_n f_n(x)$$

其中，

$$0<P_n<1,\quad \sum_n P_n=1$$

以概率 P_n，由密度函数 $f_n(x)$中抽样随机变量 η_n，它们的总体复合密度是 $f(x)$。若用 T_n 表示由 $f_n(x)$中抽样 η_n 的工作量，选取 P_n、$f_n(x)$，应使 $\sum\limits_n P_nT_n$ 尽可能的小。

例 9　用复合抽样，产生具有密度函数

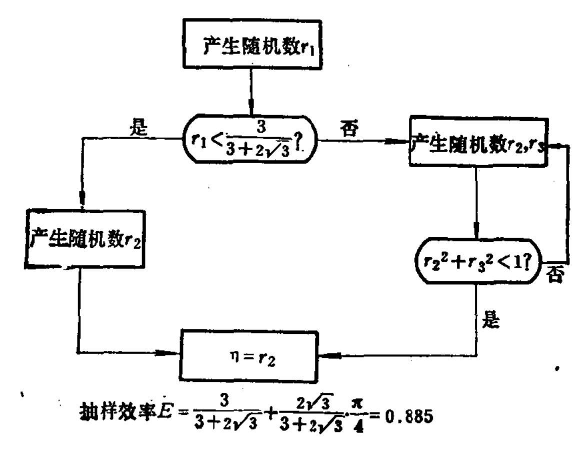
$$f(x)=\begin{cases}\dfrac{12}{(3+2\sqrt{3})\pi}\left(\dfrac{\pi}{4}+\dfrac{2\sqrt{3}}{3}\sqrt{1-x^2}\right), & 0<x<1\\ 0, & \text{其它}\end{cases}$$

的随机变量 η。

分解 $f(x)$为

$$f(x)=\frac{3}{3+2\sqrt{3}}\cdot 1+\frac{2\sqrt{3}}{3+2\sqrt{3}}\cdot\frac{4}{\pi}\sqrt{1-x^2}$$

得如图 7.9 所示的抽样算法：



抽样效率 $E=\dfrac{3}{3+2\sqrt{3}}+\dfrac{2\sqrt{3}}{3+2\sqrt{3}}\cdot\dfrac{\pi}{4}=0.885$

图 7.9　抽样框图

例 10　用复合抽样，产生 $N(0, 1)$ 分布的随机变量 u。

将正态密度函数

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

分解为四个密度函数的概率和，抽样方法如下：

(I) 以概率 $P_1=0.8638$, 产生随机数 r_1, r_2, r_3

$$u=2\left(r_1+r_2+r_3-\frac{3}{2}\right)$$

(II) 以概率 $P_2=0.1107$, 产生随机数 r_1, r_2

$$u=\frac{3}{2}(r_1-r_2)$$

(III) 以概率 $P_3=0.0228002039$, 产生随机数 r_1, r_2

令
$$x=6r_1-3$$
当 $r_2<g(x)$ 时，
$$u=x$$
这里

$$g(x)=\begin{cases}48.8751e^{-\frac{x^2}{2}}+6.0276(|x|-1.5)+13.2282(x^2-3), & |x|\leqslant 1 \text{ 时}\\ 48.8751e^{-\frac{x^2}{2}}+6.0276(|x|-1.5)-6.6141(3-|x|)^2, & 1<|x|\leqslant 1.5 \text{ 时}\\ 48.8751e^{-\frac{x^2}{2}}-6.6141(3-|x|)^2, & 1.5<|x|\leqslant 3 \text{ 时}\\ 0, & |x|>3 \text{ 时}\end{cases}$$

(IV) 以概率 $P_4=0.0026997961$, 产生随机数 r_1, r_2

令
$$x=2r_1-1,\quad y=2r_2-1$$
当 $x^2+y^2=d<1$ 时，

计算

$$u_1=x\left[\frac{9-2\ln d}{d}\right]^{1/2}$$

$$u_2=y\left[\frac{9-2\ln d}{d}\right]^{1/2}$$

若 $|u_1|>3$, 取 $u=u_1$; 否则，若 $|u_2|>3$, 取 $u=u_2$。当 $|u_1|<3$, $|u_2|<3$ 时，进行新的抽样。

§7.4　随机向量抽样

在蒙特卡洛方法的实际应用中，经常遇到多维随机向量的抽样问题。如果随机向量的各个分量相互独立，则可用 §7.3 讨论的方法，对各个分量分别独立进行抽样。实际模拟中，随机向量的各个分量多是统计相关的，抽样过程稍复杂些。这里，讨论随机向量的一般抽样方法和正态随机向量的实际抽样算法。

7.4.1　一般抽样方法

给出 n 维随机向量 $\boldsymbol{\eta}=(\eta_1, \eta_2, \cdots, \eta_n)^T$ 的联合密度函数

$$f(x_1, x_2, \cdots, x_n)$$

据此，产生 η 的抽样值。

(一)条件密度法

若随机向量 η 的密度函数 $f(x_1, x_2, \cdots, x_n)$ 可表成一串密度函数的乘积，即

$$f(x_1, x_2, \cdots, x_n)=f_1(x_1)f_2(x_2|x_1)f_3(x_3|x_1, x_2)\cdots f_n(x_n|x_1, \cdots, x_{n-1})$$

则可用 §7.3 讨论的方法进行抽样。

由密度函数 $f_1(x_1)$，产生 η_1 的抽样值 x_1。

在 $\eta_1=x_1$ 给定后，由条件密度函数 $f_2(x_2|x_1)$ 产生 η_2 的抽样值 x_2。

在 $\eta_1=x_1$, $\eta_2=x_2$ 给定后，由条件密度函数 $f_3(x_3|x_1, x_2)$，产生 η_3 的抽样值 x_3。

⋮

在 $\eta_1=x_1$, $\eta_2=x_2$, $\cdots$, $\eta_{n-1}=x_{n-1}$ 给定后，由条件密度函数 $f_n(x_n|x_1, x_2, \cdots, x_{n-1})$，产生 η_n 的抽样值 x_n。

最后，得到随机向量 η 的抽样值 $(x_1, x_2, \cdots, x_n)^T$。

以三维为例，说明由联合密度函数 $f(x, y, z)$ 计算它的条件密度函数的方法。向高维推广时，方法是类似的。

$$\left.\begin{aligned} f_1(x)&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x, y, z)\,dy\,dz \\ f_2(y|x)&=\int_{-\infty}^{+\infty}f(x, y, z)\,dz/f_1(x) \\ f_3(z|x, y)&=f(x, y, z)/f_1(x)f_2(y|x) \end{aligned}\right\} \tag{7.4.1}$$

例 11　随机向量 $(\theta, \varphi)^T$ 具有联合密度函数

$$f(\theta, \varphi)=\begin{cases}\dfrac{1}{C}\sin^2\theta\sin\varphi(1+\sqrt{3}\sin\theta\sin\varphi), & 0\leqslant\theta,\ \varphi\leqslant\dfrac{\pi}{2}\\ 0, & \text{其它}\end{cases}$$

产生它们的方向余弦

$$\begin{cases}\xi=\cos\varphi\\ \eta=\cos\theta\end{cases}$$

根据(7.4.1)，得 ξ 的边缘分布

$$f_1(x)=\begin{cases}\dfrac{12}{(3+2\sqrt{3})\pi}\left(\dfrac{\pi}{4}+2\sqrt{3}\sqrt{1-x^2}/3\right), & 0\leqslant x\leqslant 1\\ 0, & \text{其它}\end{cases}$$

抽样方法见例 9。在 $a=\sin\varphi=\sqrt{1-\xi^2}$ 给定后，η 的条件密度函数

$$f_2(y|a)=\begin{cases}\dfrac{1}{\dfrac{\pi}{4}+\dfrac{2\sqrt{3}}{3}a}\sqrt{1-y^2}(1+\sqrt{3}a\sqrt{1-y^2}), & 0\leqslant y\leqslant 1\\ 0, & \text{其它}\end{cases}$$

可用复合抽样方法进行，如图 7.10 所示。

(二)舍选法

若随机向量 η 的密度函数 $f(x_1, x_2, \cdots, x_n)$，定义在平行多面体

$$\{a_1\leqslant x_1\leqslant b_1,\ a_2\leqslant x_2\leqslant b_2,\ \cdots,\ a_n\leqslant x_n\leqslant b_n\}$$

上，且

$$\sup f(x_1, x_2, \cdots, x_n)=L$$

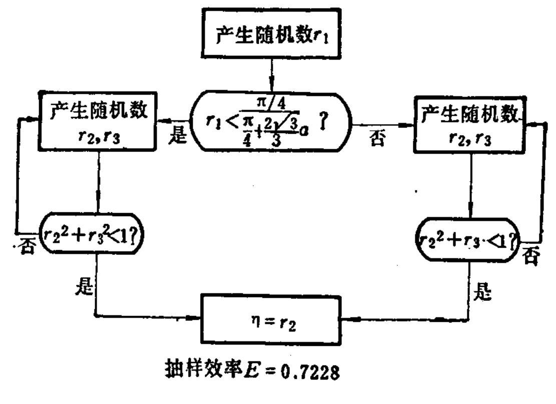


图7.10　抽样框图

取有限值。推广§7.3中的舍选法，得抽样方法：

产生随机数 $r_1, r_2, \cdots, r_n, r_{n+1}$，若

$$Lr_{n+1}<f((b_1-a_1)r_1+a_1,\ (b_2-a_2)r_2+a_2,\ \cdots,\ (b_n-a_n)r_n+a_n)$$

成立，得 η 的抽样值

$$\eta_i=(b_i-a_i)r_i+a_i\quad(i=1,\ 2,\ \cdots,\ n)$$

抽样效率

$$E=\frac{1}{L\prod_{i=1}^{n}(b_i-a_i)}$$

随维数 n 的增加而降低。

7.4.2 正态向量抽样

若随机向量 η 以函数

$$f(x_1,\ x_2,\ \cdots,\ x_n)=(2\pi)^{-\frac{n}{2}}|M|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^TM^{-1}(\boldsymbol{x}-\boldsymbol{a})\right\}\qquad(7.4.2)$$

为密度，则称 η 是 n 维正态的。这里

$$\boldsymbol{a}=\boldsymbol{E}(\boldsymbol{\eta})$$

是随机向量的数学期望，$\boldsymbol{M}$ 是 η 的 n 阶协方差矩阵，即

$$\boldsymbol{M}=\begin{pmatrix}\lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n}\\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n}\\ \cdot & \cdot & \cdots & \cdot\\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nn}\end{pmatrix}$$

其中
$$\lambda_{ij}=E[(\eta_i-a_i)(\eta_j-a_j)]=\lambda_{ji}$$

$|\boldsymbol{M}|$ 是正定矩阵 $\boldsymbol{M}$ 的行列式。

当 $\boldsymbol{a}=0$，$\boldsymbol{M}$ 为 n 阶单位矩阵时，随机向量 η 的各个分量相互独立，服从 $N(0,\ 1)$ 分布。这时，可用例3、例7、例8、例10中的方法进行抽样，并用 $\boldsymbol{u}$ 表示。一般情况下，η 和 $\boldsymbol{u}$ 存在变换关系：

$$\boldsymbol{\eta}=\boldsymbol{A}\boldsymbol{u}+\boldsymbol{a}\qquad(7.4.3)$$

线性变换矩阵 $\boldsymbol{A}$, 可用不同的方法, 由协方差矩阵 $\boldsymbol{M}$ 给出[3,15]。

取 $\boldsymbol{A}$ 为三角阵, 即

$$\boldsymbol{A}=\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix},$$

根据 $\boldsymbol{A}\boldsymbol{A}^T=\boldsymbol{M}$, 元素 a_{ij} 可按列依次算出。

对第一列诸元素有:

$$a_{11}=\sqrt{\lambda_{11}},\ a_{i1}=\lambda_{i1}/a_{11}\quad (i=2,\ 3,\ \cdots,\ n)$$

算出 1, 2, …, $j-1$ 各列元素后, 第 j 列的主对角元素

$$a_{jj}=\left[\lambda_{jj}-\sum_{k=1}^{j-1} a_{jk}^2\right]^{1/2}$$

当 $j<n$ 时, 主对角线以下各元素

$$a_{ij}=a_{jj}^{-1}\left[\lambda_{ij}-\sum_{k=1}^{j-1} a_{ik}a_{jk}\right]\quad (i=j+1,\ \cdots,\ n)$$

例 12　模拟数学期望为 0, 协方差矩阵

$$\boldsymbol{M}=\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$$

的二维正态随机向量 $\boldsymbol{\eta}$。

根据协方差矩阵 $\boldsymbol{M}$, 计算三角阵 $\boldsymbol{A}$ 的元素 a_{ij}, 有:

$$a_{11}=\sqrt{\lambda_{11}},\qquad a_{12}=0$$
$$a_{21}=a_{11}^{-1}\lambda_{21}=\lambda_{21}/\sqrt{\lambda_{11}},\quad a_{22}=[\lambda_{22}-a_{21}^2]^{1/2}=\left[\frac{\lambda_{11}\lambda_{22}-\lambda_{12}^2}{\lambda_{11}}\right]^{1/2}$$

给出正态独立向量 $\boldsymbol{u}$, 得 $\boldsymbol{\eta}$ 的抽样公式:

$$\begin{cases} \eta_1=\sqrt{\lambda_{11}}u_1 \\ \eta_2=\dfrac{1}{\sqrt{\lambda_{11}}}[\lambda_{21}u_1+(\lambda_{11}\lambda_{22}-\lambda_{12}^2)^{1/2}u_2] \end{cases}$$

§7.5　随机过程模拟

一个 n 维随机过程

$$\boldsymbol{\eta}(t)=(\eta_1(t),\ \eta_2(t),\ \cdots,\ \eta_n(t))^T$$

对固定的参量 t, 是一个 n 维随机向量。今后, 我们把参量 t 叫做时间。

用蒙特卡洛方法模拟一个系统时, 如模拟具有相关噪声的信号传输系统, 时间序列, 求解随机微分方程, 检验随机信号的滤波和预报效果等, 都要用到随机过程 $\boldsymbol{\eta}(t)$ 在离散时刻 $t_1<t_2<\cdots<t_m$ 上取值的随机序列 $\boldsymbol{\eta}(t_i)$。

在 m 个离散点上模拟随机过程 $\boldsymbol{\eta}(t)$ 就是抽样一个 $m\times n$ 维的随机向量。当 m、n 比较大时, 利用 §7.4 给出的抽样方法是相当困难的。

下面, 我们讨论几种常用随机过程的模拟方法[3,6,15]。

7.5.1 正态马尔科夫过程的模拟

设 $\boldsymbol{\eta}(t)$ 是一个 n 维正态马尔科夫过程，条件分布函数 $(s<t)$

$$F\{\eta_1(t)<x_1(t),\ \cdots,\ \eta_n(t)<x_n(t)\,|\,\eta_1(s)=x_1(s),\ \cdots,\ \eta_n(s)=x_n(s)\}$$

可由 n 维正态密度函数(7.4.2)给出。(7.4.2)中的数学期望和协方差矩阵分别由条件期望

$$\begin{aligned}\boldsymbol{a}(t|\boldsymbol{x}(s))&=\boldsymbol{E}\{\boldsymbol{\eta}(t)\,|\,\boldsymbol{\eta}(s)=\boldsymbol{x}(s)\}\\&=\boldsymbol{a}(t)+\boldsymbol{M}(t,\ s)[\boldsymbol{M}(s,\ s)]^{-1}(\boldsymbol{x}(s)-\boldsymbol{a}(s))\end{aligned}\tag{7.5.1}$$

和条件协方差矩阵

$$\boldsymbol{M}(t|\boldsymbol{x}(s))=\boldsymbol{M}(t,\ t)-\boldsymbol{M}(t,\ s)[\boldsymbol{M}(s,\ s)]^{-1}\boldsymbol{M}(s,\ t)\tag{7.5.2}$$

给出。这里，

$$\boldsymbol{a}(t)=(a_1(t),\ a_2(t),\ \cdots,\ a_n(t))^T$$

是 $\boldsymbol{\eta}(t)$ 的无条件期望，

$$\boldsymbol{M}(t;\ s)=\begin{pmatrix}\lambda_{11}(t,\ s) & \lambda_{12}(t,\ s) & \cdots & \lambda_{1n}(t,\ s)\\ \lambda_{21}(t,\ s) & \lambda_{22}(t,\ s) & \cdots & \lambda_{2n}(t,\ s)\\ \cdots & \cdots & \cdots & \cdots\\ \lambda_{n1}(t,\ s) & \lambda_{n2}(t,\ s) & \cdots & \lambda_{nn}(t,\ s)\end{pmatrix}$$

是随机过程 $\boldsymbol{\eta}(t)$ 的协方差矩阵，它的元素

$$\lambda_{ij}(t,\ s)=E[(\eta_i(t)-a_i(t))(\eta_j(s)-a_j(s))]$$

是 $\eta_i(t)$ 和 $\eta_j(s)$ 的协方差函数。

由(7.5.1)、(7.5.2)，可用多元正态随机向量的抽样方法，模拟正态马尔科夫过程。

例 13　模拟 $\boldsymbol{a}(t)=0$，协方差矩阵

$$\boldsymbol{M}(t,\ s)=\begin{pmatrix}2e^{-\beta|t-s|} & e^{-\beta|t-s|}\\ e^{-\beta|t-s|} & e^{-\beta|t-s|}\end{pmatrix}$$

的二维正态马尔科夫过程 $(\beta>0)$。

根据(7.5.1)、(7.5.2)，得

$$\boldsymbol{a}(t|\boldsymbol{x}(s))=e^{-\beta|t-s|}\begin{pmatrix}x_1(s)\\ x_2(s)\end{pmatrix}$$

$$\boldsymbol{M}(t|\boldsymbol{x}(s))=(1-e^{-2\beta|t-s|})\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}$$

由例12，计算变换矩阵 (a_{ij})，给出 $\boldsymbol{\eta}(t)$ 的模拟算法。

产生相互独立 $N(0,\ 1)$ 分布的随机向量 $\begin{pmatrix}u_{11}\\ u_{12}\end{pmatrix}$，$\begin{pmatrix}u_{21}\\ u_{22}\end{pmatrix}$，…，在 $t_1<t_2<\cdots$ 上模拟随机过程 $\boldsymbol{\eta}(t)$，得到不等时间间隔上采样的随机序列：

$$\begin{pmatrix}\eta_1(t_1)\\ \eta_2(t_1)\end{pmatrix}=\begin{pmatrix}\sqrt{2} & 0\\ \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}u_{11}\\ u_{12}\end{pmatrix}$$

$$\begin{pmatrix}\eta_1(t_i)\\ \eta_2(t_i)\end{pmatrix}=e^{-\beta|t_i-t_{i-1}|}\begin{pmatrix}\eta_1(t_{i-1})\\ \eta_2(t_{i-1})\end{pmatrix}+\left[\frac{1-2e^{-2\beta|t_i-t_{i-1}|}}{2}\right]^{1/2}\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}u_{1i}\\ v_{2i}\end{pmatrix}$$

7.5.2 有理谱正态平稳过程的模拟

设一维正态平稳过程 $\eta(t)$，具有数学期望

$$E[\eta(t)]=0$$

协方差函数

$$\lambda(\tau)=E[\eta(t)\cdot\eta(t+\tau)]=\sum_i(a_i\cos\omega_i\tau+b_i\sin\omega_i|\tau|)e^{-\beta_i|\tau|}$$

其中，常量 $\omega_i>0$, $\beta_i>0$, 系数 a_i, b_i 是 τ 的偶次多项式。

随机过程 $\eta(t)$ 的谱密度

$$S_\eta(\omega)=\int_{-\infty}^{+\infty}\lambda(\tau)e^{-j\omega\tau}d\tau=\frac{R_0+R_1\omega^2+\cdots+R_k\omega^{2k}}{L_0+L_1\omega^2+\cdots+L_l\omega^{2l}}=\frac{R(\omega)}{L(\omega)} \tag{7.5.3}$$

是 ω 的有理函数，系数 R_p, L_q 是实数($p=0, 1, \cdots, k$; $q=0, 1, \cdots, l$; $k<l$)。

在物理上，随机过程 $\eta(t)$ 可以看作是一个常系数线性系统以白噪声为输入时的输出信号。设 $k(j\omega)$ 为系统的传递函数，则 $S_\eta(\omega)$ 和白噪声的谱密度 $S_u(\omega)$ 存在关系

$$S_\eta(\omega)=|k(j\omega)|^2S_u(\omega) \tag{7.5.4}$$

白噪声的谱密度 $S_u(\omega)$ 可取为 1，由此得到线性系统的传递函数。

对平稳过程 $\eta(t)$，求谱密度 $S_\eta(\omega)$ 两个多项式

$$R(\omega)=R_0-R_1(j\omega)^2+\cdots+(-1)^kR_k(j\omega)^{2k}=0$$

$$L(\omega)=L_0-L_1(j\omega)^2+\cdots+(-1)^kL_l(j\omega)^{2l}=0$$

的复根 $j\omega R_p$, $j\omega L_q$。由系统的稳定性，舍去左半平面上的复根，得

$$S_\eta(\omega)=\frac{R_0}{L_0}\left|\frac{(j\omega-j\omega R_1)\cdots(j\omega-j\omega R_k)}{(j\omega-j\omega L_1)\cdots(j\omega-j\omega L_l)}\right|^2$$

$$=\left|\frac{\rho_0+\rho_1(j\omega)+\cdots+\rho_k(j\omega)^k}{v_0+v_1(j\omega)+\cdots+v_l(j\omega)^l}\right|^2=|k(j\omega)|^2$$

于是得到常系数线性系统的传递函数

$$k(j\omega)=\frac{\rho_0+\rho_1(j\omega)+\cdots+\rho_k(j\omega)^k}{v_0+v_1(j\omega)+\cdots+v_l(j\omega)^l} \tag{7.5.5}$$

根据(7.5.4)、(7.5.5)，得到随机过程 $\eta(t)$ 的模拟算法：

(1) 产生具有谱密度

$$S_\xi(\omega)=\frac{1}{L(\omega)}=\frac{1}{L_0+L_1\omega^2+\cdots+L_l\omega^{2l}}$$

的正态平稳过程 $\xi(t)$ 及其导数 $\xi'(t)$, $\xi''(t)$, $\cdots$, $\xi^{(l-1)}(t)$。$\xi(t)$ 具有协方差函数

$$\lambda_\xi(\tau)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}s_\xi(\tau)e^{j\omega\tau}d\tau$$

为了模拟 $\xi(t)$ 及其各阶导数 $\xi^{(i)}(t)$ ($i=1, 2, \cdots, l-1$)，可用模拟一个 l 维正态马尔科夫过程

$$\boldsymbol{\xi}(t)=(\xi(t), \xi'(t), \cdots, \xi^{(l-1)}(t))^T$$

来实现。它的无条件期望 $E(\boldsymbol{\xi}(t))=0$，协方差函数

$$\lambda_{qp}(t-s)=E[\xi^{(q)}(t)\xi^{(p)}(s)]=\frac{\partial^{q+p}\lambda_\xi(t-s)}{\partial t^q\partial s^p}$$

(2) 产生 $\eta(t)$

$$\eta(t)=\rho_0\xi(t)+\rho_1\xi'(t)+\cdots+\rho_k\xi^{(k)}(t)$$

例 14　模拟 $E[\eta(t)]=0$, $\lambda_\eta(\tau)=\sigma^2 e^{-\beta|\tau|}(\beta>0)$的正态平稳过程 $\eta(t)$。

根据(7.5.3)、(7.5.4),得到平稳过程 $\eta(t)$的谱密度

$$S_\eta(\omega)=\int_{-\infty}^{+\infty}\sigma^2 e^{-\beta|\tau|}e^{-j\omega\tau}d\tau=\frac{2\beta\sigma^2}{\beta^2+\omega^2}$$

和线性系统的传递函数

$$k(j\omega)=\frac{\sigma\sqrt{2\beta}}{\beta+j\omega}$$

给出 $\eta(t)$的模拟公式:

产生 $N(0,\ 1)$的简单子样 $u_1,\ u_2,\ \cdots$,

$$\eta(t_1)=\sigma\sqrt{2\beta}u_1$$
$$\eta(t_i)=e^{-\beta|t_i-t_{i-1}|}\eta(t_{i-1})+\sigma\sqrt{1-e^{-2\beta|t_i-t_{i-1}|}}\,u_i$$

实际模拟时,为减少偏离,可把 u_1 取为 $\eta(t)$的数学期望值,即令 $u_1=0$。

例 15　模拟均值为 0, 协方差函数

$$\lambda(\tau)=\sigma^2\cos\omega\tau e^{-\beta|\tau|}$$

的正态平稳过程 $\varphi(t)$。

设 $\eta_1(t)$、$\eta_2(t)$是两个相互独立、以 $\sigma^2 e^{-\beta|\tau|}$ 为协方差函数的正态平稳过程(模拟算法见例 14), 则有

$$\varphi(t)=\eta_1(t)\cos\omega t+\eta_2(t)\sin\omega t$$

7.5.3　非平稳过程的模拟

在许多情况下,非平稳正态过程可由正态平稳过程经变换得到。

设 $\eta(t)$是一个均值为 0, 协方差函数为 $\lambda_\eta(\tau)$的正态平稳过程,

$$\varphi(t)=f(t)\eta(t)+g(t) \tag{7.5.6}$$

当 $f(t)$, $g(t)$是时间 t 的确定性函数时,是一个非平稳的随机过程,且有

$$E[\varphi(t)]=g(t)$$
$$\lambda_\varphi(t,\ s)=f(t)f(s)\lambda_\eta(t-s)$$

利用(7.5.6),可以给出很大一类非平稳过程的渐近模拟算法。

§7.6　随机数的检验

上面几节讨论了随机数、随机变量、随机过程的模拟方法。很自然的要问,用这些方法产生的数值序列,具有我们所要求的统计性质么?能够在蒙特卡洛模拟中,放心大胆地使用么?这里,我们从统计假设检验出发,分析它们的统计性质,讨论和解决上述问题。

设随机变量 η 具有连续的分布函数 $F(x)$,则随机变量

$$R=F(\eta) \tag{7.6.1}$$

均匀分布在(0, 1)上。因此,能对均匀分布的随机数 R 进行统计检验, 通过 $R=F(\eta)$, 也就可以检验以 $F(x)$为分布函数的随机变量 η。均匀分布的简单性, 也便于设计一些简单有效的统计检验方法。所以,这里只讨论均匀分布随机数的检验问题。

随机数的统计检验就是根据(0, 1)上均匀总体简单子样$\{R_i\}$的性质,如分布参数, 均匀性、独立性、随机性和组合规律性等, 研究我们产生的随机数序列 $\{r_i\}$ 的相应性质,进行比

较，视其差异显著与否，决定取舍。如各类统计检验的差异并不显著，则可接受$\{r_i\}$为均匀总体R的简单子样。

在介绍随机数的各类具体检验方法之前，先给出统计检验中两个常用统计量的构造和检验方法。

(1) 设随机变量η具有数学期望$E(\eta)=a$和有限的方差$D(\eta)=\sigma^2$。在N次独立试验中，得到η的N个观测值

$$\eta_1,\ \eta_2,\ \cdots,\ \eta_N$$

当N充分大时，根据中心极限定理，统计量

$$u=\left(\frac{1}{N}\sum_{i=1}^{N}\eta_i-a\right)\Big/\frac{\sigma}{\sqrt{N}} \tag{7.6.2}$$

以$N(0,\ 1)$为极限分布。

取显著水平$\alpha=0.05$，根据概率论中的实际推论原理，当$|u|>1.96$时，称差异显著，可拒绝$E(\eta)=a$的假设。

(2) 将随机变量η的简单子样$\eta_1,\ \eta_2,\ \cdots,\ \eta_N$，按一定规则，分为互不相交的$k$组。记第$i$组的观测频数为$n_i$。若随机变量$\eta$属于第$i$组的概率为$p_i$，有理论频数$m_i=Np_i(i=1,\ 2,\ \cdots,\ k)$，统计量

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-m_i)^2}{m_i} \tag{7.6.3}$$

渐近服从自由度$k-l-1$的χ^2-分布，简记为$\chi^2(k-l-1)$。这里，l是确定p_i时，由子样η_i中给出的约束条件数。

统计检验时，取显著水平$\alpha=0.05$，由χ^2-分布表中可查到$\chi^2_{0.05}$。当$\chi^2>\chi^2_{0.05}$时，差异显著，拒绝假设$m_i=Np_i$。

为了更有效地进行统计检验，一般要求：

在(7.6.2)中，子样数$N>30$；

在(7.6.3)中，$k\geqslant5$，$m_i\geqslant5$，且k和N存在渐近最优关系

$$k=1.87(N-1)^{2/5}$$

由此，可得N-k对照表：

N	500	1,000	2,000	5,000	10,000	50,000
k	22	30	39	56	74	142

当χ^2统计量的自由度$d>30$时，

$$u=\sqrt{2\chi^2}-\sqrt{2d-1}$$

服从$N(0,\ 1)$分布。

设

$$r_1,\ r_2,\cdots,\ r_N \tag{7.6.4}$$

是需要进行统计检验的一组随机数。根据(7.6.2)、(7.6.3)，构造一些具体的检验方法。

7.6.1 参数检验

随机数的参数检验是检验它的分布参数的观测值和理论值的差异是否显著。

由观测序列(7.6.4)，得到随机数的一阶矩、二阶矩和方差的观测值：

$$\bar{r}=\frac{1}{N}\sum_{i=1}^{N}r_i$$

$$\bar{r^2}=\frac{1}{N}\sum_{i=1}^{N}r_i^2$$

$$s^2=\frac{1}{N}\sum_{i=1}^{N}\left(r_i-\frac{1}{2}\right)^2=\bar{r^2}-\bar{r}+\frac{1}{4}$$

根据随机数的理论分布，不难计算

$$E(\bar{r})=\frac{1}{2}\qquad D(\bar{r})=\frac{1}{12N}$$

$$E(\bar{r^2})=\frac{1}{3}\qquad D(\bar{r^2})=\frac{4}{45N}$$

$$E(s^2)=\frac{1}{12}\qquad D(s^2)=\frac{1}{180N}$$

由(7.6.2)，得

$$u_1=\sqrt{12N}\left(\bar{r}-\frac{1}{2}\right)$$

$$u_2=\frac{1}{2}\sqrt{45N}\left(\bar{r^2}-\frac{1}{3}\right)$$

$$u_3=\sqrt{180N}\left(s^2-\frac{1}{12}\right)$$

渐近服从 $N(0, 1)$ 分布。

7.6.2 均匀性检验

随机数的均匀性检验，又称频率检验，检验它的经验频率和理论频率的差异是否显著。

(一)拟合均匀性检验

把(0, 1)区间分为 k 个等区间。按 r_i 取值的大小，分为 k 类。设有 n_j 个随机数属于第 j 类，即共有 n_j 个 r_i 满足

$$\frac{j-1}{k}\leqslant r_i<\frac{j}{k}\quad (j=1, 2, \cdots, k)$$

这里

$$\sum_{j=1}^{k}n_j=N$$

根据均匀性假设，r_i 落在每个小区间的概率相等，即

$$P_j=\frac{1}{k}$$

有理论频数

$$m_j=N/k\quad (j=1, 2, \cdots, k)$$

根据(7.6.3)，得统计量

$$\chi^2=\frac{k}{N}\sum_{j=1}^{k}\left(n_j-\frac{N}{k}\right)^2$$

渐近服从自由度 $k-1$ 的 χ^2-分布。统计量

$$\sqrt{N}D_N=\frac{1}{\sqrt{N}}\max_{1<j<k}\left|\sum_{i=1}^{j}\left(n_i-\frac{N}{k}\right)\right|$$

为累积频率检验，渐近服从柯尔莫果洛夫-斯米尔诺夫 λ-分布。取显著水平 $\alpha=0.05$，当 $\sqrt{N}D_N>1.35$ 时，拒绝原假设。

(二)分位均匀性检验

在二进制数字计算机上，每一个随机数 r_i 都是由若干位 0 或 1 排列组合而成的。根据要求，各位数应相互独立，以概率 $\frac{1}{2}$ 取值 0 或 1。取 r 的前 l 位二进制数，它的数字和，即 l 位中出现 1 的个数，是进行 l 次独立试验时的成功次数。对 N 个随机数(7.6.4)取前 l 位数进行试验，得到不成功，成功 1 次，成功 2 次，……，成功 l 次的观测频数 $n_0, n_1, \cdots, n_l$。在 l 次独立试验中，j 次成功的概率 P_j 服从二项分布

$$P_j=C_l^j\left(\frac{1}{2}\right)^l \quad (j=0,\ 1,\ \cdots,\ l)$$

给出理论频数 $m_j=NP_j$，得到构造统计量(7.6.3)的假设。

取 $l=10$ 时，得 P_j

j	0	1	2	3	4	5	6	7	8	9	10
P_j	$\frac{1}{1024}$	$\frac{10}{1024}$	$\frac{45}{1024}$	$\frac{120}{1024}$	$\frac{210}{1024}$	$\frac{252}{1024}$	$\frac{210}{1024}$	$\frac{120}{1024}$	$\frac{45}{1024}$	$\frac{10}{1024}$	$\frac{1}{1024}$

由 P_j 表可以看出，为了有效的利用检验(7.6.3)，必须将两端的概率 P 进行合并以增大理论频数。

7.6.3 独立性检验

随机数的独立性检验，重点检验(7.6.4)中前后各数的统计相关是否异常。

(一)相关系数检验

相关系数取值为 0 是两个随机变量相互独立的必要条件，取值大小，给出它们之间线性相关强弱的测度，故可用来检验随机数的独立性。

计算(7.6.4)前后距离为 j 的相关系数

$$\bar{\rho}_j=\left[\frac{1}{N-j}\sum_{i=1}^{N-j}r_ir_{j+i}-(\bar{r})^2\right]\Big/s^2$$

对充分大的 $N(N-j>50)$，取零假设 H_0: $\rho=0$，统计量

$$u=\bar{\rho}_j\sqrt{N-j}$$

渐近服从 $N(0,\ 1)$ 分布。

(二)联列表检验

在$(X\text{-}Y)$平面上，将单位正方形分为 k^2 个相等的小正方形。将随机数序列(7.6.4)，按出现的先后顺序两两分组，如取

$$(r_1,\ r_{l+1}),\ (r_2,\ r_{l+2}),\ \cdots,\ (r_N,\ r_l)$$

进行试验$(l=1,\ 2,\ \cdots)$。记落入小正方形$(i,\ j)$内的观测频数为 $n_{ij}(i,\ j=1,\ 2,\ \cdots,\ k)$，令

$$n_{i\cdot}=\sum_{j=1}^{k}n_{ij}, \qquad n_{\cdot j}=\sum_{i=1}^{k}n_{ij}$$

根据独立性假设，统计量

$$\chi^2=N\left\{\sum_{i,j=1}^{k}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right\}$$

渐近服从分布 $\chi^2[(k-1)^2]$。

7.6.4 组合规律性检验

随机数的组合规律性检验是按照随机数出现的先后顺序，根据一定的规则进行组合，检验组合观测值和理论值的差异是否显著。

(一)距离检验

一对随机数(r_1, r_2)组成 X-Y 平面上单位正方形内的一个随机点。两个随机点(r_1, r_2)，(r_3, r_4)之间的距离 d，是由四个随机数组成的一个随机变量，可用来检验随机数。

取
$$d^2=(r_1-r_3)^2+(r_2-r_4)^2$$
有分布函数

$$F(\alpha^2)=\begin{cases}0, & \alpha^2\leqslant 0\\ \pi\alpha^2-\dfrac{8}{3}\alpha^3+\dfrac{1}{2}\alpha^4, & 0<\alpha^2\leqslant 1\\ \dfrac{1}{3}+(\pi-2)\alpha^2+4(\alpha^2-1)^{1/2}+\dfrac{8}{3}(\alpha^2-1)^{3/2}-\dfrac{\alpha^4}{2} & \\ \quad -4\alpha^2\operatorname{arc\,csc}\alpha, & 1<\alpha^2\leqslant 2\\ 1, & \alpha^2>2\end{cases}$$

把(0, 2)区间分为一些不相交的子区间，由 $F(\alpha^2)$给出理论概率，求出 d^2 的理论频数 m_j。根据(7.6.3)，即可构造对(7.6.4)进行统计检验的统计量。

(二)配套检验

从随机数序列(7.6.4)的第一个随机数开始，取其第一个数字(如取 k 进制数，$k=8$, 10 或 16)记下来，略去那些已经出现过的一些数字，直到用 l 个随机数配齐全部 k 个不同数字 0, 1, …, $k-1$ 为止。随机变量 l 构成我们进行检验的统计量。

对 N 个随机数进行试验，得到配齐一套数字用到 l 个随机数的观测频数 $n_l(l=k, k+1, \cdots)$。若$\{r_i\}$相互独立，均匀分布在(0, 1)上，则随机变量 l 具有概率

$$P_k(l)=\sum_{1}^{k}(-1)^i C_k^i\left(1-\frac{i}{k}\right)^l \quad (l=k,\ k+1,\ \cdots)$$

且

$$E(l)=k\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{k-1}+\frac{1}{k}\right)$$
$$D(l)=k\left(\frac{1}{(k-1)^2}+\frac{2}{(k-2)^2}+\cdots+\frac{k-2}{2^2}+\frac{k-1}{1^2}\right)$$

由此，可构造对 l 进行统计检验的统计量 u 和 χ^2。

如在 $k=16$ 时，$E(l)=54.09$, $D(l)=(18.75)^2$

配齐一套数字用到 $l\sim m$ 个随机数的概率 P 为：

$l\sim m$	概率 P	$l\sim m$	概率 P	$l\sim m$	概率 P	$l\sim m$	概率 P
16～30	0.04629	40～42	0.07659	52～54	0.06720	66～70	0.05705
31～33	0.04340	43～45	0.07884	55～57	0.06039	71～75	0.04324
34～36	0.05838	46～48	0.07735	58～60	0.05323	76～90	0.07388
37～39	0.06980	49～51	0.07313	61～65	0.07392	90 以上	0.04731

7.6.5 连检验

把随机序列按一定规则进行分类。如分为两类,记为 a、b, 得到形如 $aabbabbbaaaba\cdots$ 由两类元素组成的随机序列。我们把夹于异类元素之间的同类元素, 如 $\cdots abbba\cdots$ 中的 b 类元素, 称为一个连(或游程)[14], 所含同类元素的个数称为连长。出现连长为 i 的连数记为 l_i, $l=\sum_i l_i$ 称为总连数,组成进行统计检验的随机量。因此, 随机数的连检验, 按随机数出现的先后顺序,重点检验它的连贯现象是否异常。

(一)正负连检验

把随机数序列$\left\{r_i-\dfrac{1}{2}\right\}$按正负分为两类。根据均匀性、独立性假设,出现 a、b 两类元素的概率都是 1/2, 且有

$$E(l)=\frac{N}{2}+1,\quad D(l)=\frac{N-1}{4}$$

$$P\{l=k\}=\frac{1}{2^k}\quad(k=1,\ 2,\ \cdots)$$

给出按照(7.6.2)、(7.6.3)构造统计量 u、χ^2 的假设。

(二)升降连检验

把随机序列$\{r_i-r_{i-1}\}$, 按照正负分为两类连,表示随机数的增减及其长度的变化规律, 称为升降连。这时有

$$E(l)=\frac{2N-1}{3},\quad D(l)=\frac{16N-29}{90}$$

$$P_1=5/8,\ P_2=11/40,\ P_3=19/240,\ P_4=29/1680,\ \cdots$$

给出按照(7.6.2)、(7.6.3)构造统计量 u、χ^2 的假设。

下面,作为一个例子,给出对乘同余法产生随机数进行统计检验的一些结果。

用乘同余法

$$x_{n+1}\equiv\lambda x_n\quad(\mathrm{mod}\ 2^{31})$$

产生伪随机数。取参数

$$\lambda=5^{13}=9184472\bar{0}\quad x_0=19720712$$

这里, λ、x_0 为十六进制数,最后一位有一虚设的二进位。

$$r_n=2^{-31}x_n,\quad n=1,\ 2,\ \cdots,\ N$$

取 $N=500$, 进行统计检验,得如下结果。

(1) 参数检验

最小值 $\min\{r_1,\ r_2,\ \cdots,\ r_N\}=0.000026$

最大值 $\max\{r_1,\ r_2,\ \cdots,\ r_N\}=0.999989$

均值 $\bar{r}=0.499221$

标准差 $s=0.293373$

(2) 均匀性检验 将(0, 1)区间分为 20 个等区间,得检验值

$$\chi^2(19)=10.56,\quad \sqrt{N}D_N=0.3578$$

(3) 独立性检验

(i) 相关系数检验

$\bar{\rho}_1=-0.042604$

$\bar{\rho}_2=\ \ 0.006044$

$\bar{\rho}_3=-0.081950$

$\bar{\rho}_4=\ \ 0.062602$

(ii) 联列表检验($k=5$)

$\chi_1^2=15.46$

$\chi_2^2=20.85$

$\chi_3^2=21.34$

$\chi_4^2=15.62$

(4) 连检验

正负连检验			升降连检验		
连长	理论值 P	观测值 $\bar{P}$	连长	理论值 P	观测值 $\bar{P}$
1	0.5	0.500000	1	0.625	0.626911
2	0.25	0.257812	2	0.275	0.253822
3	0.125	0.152344	3	0.079167	0.088685
4	0.0625	0.035156	4～∞	0.020833	0.030582
5	0.03125	0.023438			
6～∞	0.03125	0.031250			
总连数	251	256	总连数	333	327
$\chi^2(5)$	10.906		$\chi^2(3)$	2.4008	
u	0.4919		u	−0.6376	

§7.7 加速收敛原理

用蒙特卡洛方法模拟一个问题的数值结果，可以看作一个高维积分的数值求积。如在模拟计算中，一次试验需要用到m个随机数r，则模拟过程是对随机数的m维函数

$$f(r_1,\ r_2,\ \cdots,\ r_m)$$

进行抽样。统计量$f(r_1,\ r_2,\ \cdots,\ r_m)$的数学期望$\theta$未知，是我们所要求的解。根据数学期望的定义

$$\theta=E[f(r_1,\ \cdots,\ r_m)]=\int_0^1\cdots\int_0^1 f(x_1,\ \cdots,\ x_m)dx_1\cdots dx_m \qquad (7.7.1)$$

因此，蒙特卡洛模拟的结果也就是积分(7.7.1)的近似值。显然，这一看法，并不总是适宜的。但是，模拟积分(7.7.1)的加速收敛原理和各种降低模拟方差的技巧，对于一般蒙特卡洛模拟有着普遍的意义。

下面，为简化叙述，以模拟一维积分

$$\theta=\int_0^1 f(x)dx \quad (0<f(x)<1) \qquad (7.7.2)$$

为例，说明加速收敛原理和降低模拟方差的一些技巧。

在§7.1中，我们给出积分(7.7.2)两种不同的模拟算法。为了比较同一概型、两种不同模拟算法的好坏，这里引入方法效率比的概念。

令T_1、T_2，σ_1^2、σ_2^2，分别表示同一概型中两种不同模拟算法在一次模拟试验里的平均运算量(或计算时间)和方差。为了达到给定的模拟精度，根据(7.1.5)、(7.1.7)，计算时间正比于乘积$T\sigma^2$。在实际应用中，可在计算机上通过次数不多的模拟试验，给出T、σ^2的渐近估计。

定义

$$C=\frac{T_2\sigma_2^2}{T_1\sigma_1^2} \tag{7.7.3}$$

为模拟算法 I 对模拟算法 II 的效率比。

现在，讨论两种常用的模拟算法(7.1.2)、(7.1.4)，分析它们的方差来源，给出加速收敛的基本原理。

在模拟算法(7.1.4)中，取随机变量

$$\eta_i^{(1)}=f(r_i)$$

统计量

$$\theta_1=\frac{1}{N}\sum_{i=1}^{N}f(r_i)$$

给出 θ 的无偏估计，一般称为期望值估计法。$\eta^{(1)}$ 的方差

$$\sigma_1^2=\int_0^1[f(x)-\theta]^2dx$$

是由于 x 跑遍整个积分区域(0, 1)时，$f(x)$ 对期望值 θ 的变异引起的。由此，可以给出第一个加速收敛原理。

在 $f(x)$ 的模拟过程中，减少 $f(x)$ 对期望值 θ 的变异，可以降低结果的误差，提高收敛速度。

在模拟算法(7.1.2)中，定义随机变量

$$\eta^{(2)}(r_1,\ r_2)=\begin{cases}1, & \text{当 } r_1<f(r_2)\text{ 时}\\ 0, & \text{当 } r_1>f(r_2)\text{ 时}\end{cases}$$

统计量

$$\theta_2=\frac{1}{N}\sum_{i=1}^{N}\eta^{(2)}(r_{2i-1},\ r_{2i})$$

给出 θ 的无偏估计，一般称为随机投点估计法，等价于从概率 $P=\theta$ 的二项分布中进行抽样，故有方差

$$\sigma_2^2=\theta(1-\theta)$$

因

$$T_2>T_1$$

$$\sigma_2^2-\sigma_1^2=\int_0^1 f(x)[1-f(x)]dx>0$$

故两种模拟算法的效率比

$$C=\frac{T_2\sigma_2^2}{T_1\sigma_1^2}>1$$

算法 θ_1 较算法 θ_2 有效。

分析随机变量 $\eta^{(2)}$，有

$$\int_0^1\eta^{(2)}(x_1,\ x_2)dx_1=f(x_2)$$

故知，由于 θ_1 的模拟中，应用 $\eta^{(2)}(r_1,\ r_2)$ 对随机变量 r_1 的数学期望 $f(r_2)$ 代替 $\eta^{(2)}(r_1,\ r_2)$，形成 $\sigma_1^2<\sigma_2^2$，降低了方差。因此，在蒙特卡洛模拟里，如果模拟过程中，在一处能用理论分析的数学期望值代替在该处的统计模拟，可以减少结果的误差。这是加速收敛的第二个基本原理。

下面，从减少模拟量在定义域上对结果 θ 的变异，尽可能应用理论分析的期望值代替模

拟估计值两个加速收敛原理出发,讨论几种降低方差的具体抽样方法[8,14,15]。

(1) 重要抽样

设 $g(x)$ 是 $(0,1)$ 上随机变量 ξ 的密度函数。变积分(7.7.2)为

$$\theta=\int_0^1 f(x)dx=\int_0^1 \frac{f(x)}{g(x)}\cdot g(x)dx$$

引入随机变量

$$\eta^{(3)}=\frac{f(\xi)}{g(\xi)}$$

其期望值为 θ, 方差

$$\sigma_3^2=\int_0^1 \frac{f^2(x)}{g(x)}dx-\theta^2 \tag{7.7.4}$$

取统计量

$$\theta_3=\frac{1}{N}\sum_{i=1}^{N}\eta_i^{(3)}$$

给出 θ 的无偏估计。

为了降低 σ_3, 从它的计算公式(7.7.4)可知,取 $f(x)/g(x)=\theta$, 即

$$g(x)=\frac{1}{\theta}f(x) \tag{7.7.5}$$

得 $\sigma_3^2=0$ 的零方差估计。显然,在实际模拟中, θ 是未知的, (7.7.5)中的取法并不具有什么实际意义。但是,从 $\eta^{(3)}$ 的抽样中可以看出, 对任意一个密度函数 $g(x)$ 进行抽样, 用 f/g 代替 f 进行补偿,都能得到 θ 的无偏估计。我们的目的是减少估计方差 σ_3^2。只要选取同 $f(x)$ 尽可能类似的函数 $g(x)$,使比值 $f(x)/g(x)$ 接近一个常量,就可以得到降低方差的算法。

根据被积函数 $f(x)$ 的性质, 由类似于 $f(x)$ 的 $g(x)$ 中进行抽样, 就是改变原来的 $(0,1)$ 上的平均抽样,在积分区域的重要部分,即对 θ 贡献大的部分,选取比较多的抽样点,故称为重要抽样。

(2) 分层抽样

分层抽样,就其想法和重要抽样一样,都是减少 $f(x)$ 对期望值 θ 的变异。

把积分区域 $(0,1)$ 分为一些不相交的子区间。 在每个小区间上, 进行均匀抽样。 样本量 n_j 正比于该区间上对积分值的贡献。 这种分层抽样, 不增加太多的计算量, 达到降低方差的目的。

(3) 相关抽样

利用随机变量之间的相关以减少模拟方差的抽样方法称为相关抽样。要利用相关抽样,必须进一步研究被积函数 $f(x)$ 的性质。

若 $f(x)$ 是 $(0,1)$ 上的单调函数, 采用对称化的处理方法, 可以减少 $f(x)$ 对积分值 θ 的变异。如取

$$g(x)=\frac{1}{2}[f(x)+f(1-x)]$$

随机变量

$$g(R)=\frac{1}{2}[f(R)+f(1-R)]$$

以 θ 为期望值。由于 $f(x)$ 的单调性,随机变量 $f(R)$、$f(1-R)$ 负相关。$g(R)$ 的方差

$$\sigma_4^2=\frac{1}{4}D[f(R)+f(1-R)]=\frac{\sigma_1^2}{2}[1+\rho]$$

这里，$\sigma_1^2=D[f(R)]$，$\rho<0$，是$f(R)$和$f(1-R)$的相关系数。一般情况下，$\sigma_4^2<\sigma_1^2$。

进一步分析$f(x)$的性质，也可以构造正相关的抽样方法。

取和$f(x)$相近、积分值θ_0已知的函数$\phi(x)$。则随机变量

$$g(R)=f(R)-\phi(R)+\theta_0$$

以θ为期望，且$f(R)$、$\phi(R)$的相关系数$\rho>0$。$g(R)$的方差

$$\sigma_5^2=D[f(R)-\phi(R)+\theta_0]=\sigma_1^2+\sigma^2[\phi(R)]-2\rho\sigma_1\sigma[\phi(R)]$$

当$\phi(x)$能够吸收被积函数$f(x)$对θ的大部分变异时，$\sigma_5^2<\sigma_1^2$成立。实际上，这里用理论分析得到$\phi(x)$的积分值θ_0代替部分模拟，降低结果的方差σ_5^2。

在文献中，也把负相关($\rho<0$)的抽样叫做对偶变数，正相关($\rho>0$)的抽样称为控制变数。

(4) 序贯抽样

上面讨论的几种加速收敛方法，各次试验相互独立。近来已把统计上的序贯抽样用于统计模拟。模拟过程中，模拟的结果不仅取决于本次试验，而且还依赖于以前各次的试验结果，以加速模拟过程的收敛。

我们用蒙特卡洛方法求解非线性代数方程组为例，说明这一抽样方法。

给出非线性代数方程组

$$\begin{cases}f_1(x_1,\ x_2,\ \cdots,\ x_n)=0\\ f_2(x_1,\ x_2,\ \cdots,\ x_n)=0\\ \cdots\cdots\\ f_n(x_1,\ x_2,\ \cdots,\ x_n)=0\end{cases}$$

求出使

$$Q=\sum_{i=1}^{n}\alpha_i^2f_i^2(x_1,\ x_2,\ \cdots,\ x_n)$$

取值最小的点$(x_1,\ x_2,\ \cdots,\ x_n)$，作为代数方程组的近似解。这里，加权值$\alpha_i$取常量。

根据方程组的特点和物理意义，一般可以定出求解区域D_0。在区域D_0上随机取点，构造点集满足不等式：

$$Q(\boldsymbol{x}_{(1)})>Q(\boldsymbol{x}_{(2)})>\cdots>Q(\boldsymbol{x}_{(k)})$$

计算点集$\boldsymbol{x}_{(1)},\ \boldsymbol{x}_{(2)},\ \cdots,\ \boldsymbol{x}_{(k)}$的均值$\boldsymbol{x}$，方差$s^2$，以$\boldsymbol{x}$为中心，$2s$为区间长度，给出新的求解区域$D$。这种压缩抽样区域，加以必要的人工干预的办法，经验表明可以提高收敛速度。

最后，我们以$f(x)=e^{x-1}$为例，计算积分(7.7.2)，说明各种不同加速收敛方法的效果。

在重要抽样中，取

$$g(x)=\frac{2}{3}(1+x)$$

正相关抽样中，取

$$\phi(x)=(1+x)/e$$

负相关抽样中，取

$$f(1-x)$$

有如下结果：

抽 样 方 法	均 方 差	σ/σ_1
期 望 估 计	0.180986	1
随机投点估计	0.482228	2.7
重 要 抽 样	0.266140	0.37
正相关抽样	0.076860	0.43
负相关抽样	0.023011	0.13

§7.8 蒙特卡洛应用

应用蒙特卡洛方法，可以解决许多类型不同的数学物理、工程技术问题。这些问题大致可以分为两类，即确定性问题和随机性问题。

用蒙特卡洛方法求解严格确定的数学物理问题，主要困难是构造一个简单、适用的概型，使这个问题的解对应于概型中随机变量的概率分布或参量。在计算机上，对给定的概型进行蒙特卡洛模拟。大量模拟试验后，给出分布或参量的统计估计值，作为问题的近似数值解。计算高维积分，求解线性、非线性代数方程组，计算逆矩阵，模拟椭圆型、抛物型方程，求解线性算子的特征值等[1,14]，都是这一类型的问题。但在实际计算中，这类方法应用并不很多。

模拟随机性的工程技术、数学物理问题是蒙特卡洛方法的主要应用领域。给出几个简单例子，说明实际模拟的过程。

(1) 回归方程的模拟检验

在回归分析计算中，对得到的回归方程，要进行预报或控制效果可靠性和稳定性的检验。一般方法是把给出的观测数据分为两部分，一部分(大量的)用来建立回归方程，一部分(少量的)用来检验回归方程。当观测数据量不大时，应用这种方法是困难的。为了克服这种困难，用蒙特卡洛方法伪造观测数据中的预报量[7]，检验回归方程的显著性。

给出 N 组观测数据

$$x_{n1},\ x_{n2},\ \cdots,\ x_{nm};\ y_n \quad n=1,\ 2,\ \cdots,\ N;\ N>m$$

假定预报量 y 满足回归方程

$$y_n=\beta_0+\sum_{i=1}^{m}\beta_i x_{ni}+\varepsilon_n \tag{7.8.1}$$

这里，ε_n 相互独立，服从 $N(0,\ \sigma^2)$ 分布。利用逐步回归算法(参见本书"回归分析"一章)舍选预报因子 x_i，得到回归方程，给出复相关系数

$$R^*=\sqrt{1-\frac{\sum\limits_{n=1}^{N}(y_n^*-\bar{y})^2}{\sum\limits_{n=1}^{N}(y_n-\bar{y})^2}}$$

这里，y_n^* 表示 y_n 的回归预报值，

$$\bar{y}=\frac{1}{N}\sum_{n=1}^{N}y_n$$

是 y 的平均值。

为了检验用回归模型(7.8.1)和随机预报的差异是否显著，用蒙特卡洛方法伪造 N 个

因变量 y。例如,用§7.3给出的方法,产生 $N(0,\ s_y^2)$ 分布的正态随机变量。这里

$$s_y^2=\frac{1}{N-1}\sum_{n=1}^{N}(y_n-\bar{y})^2$$

在预报量 y 和预报因子 x_i 无关的假定下,是方差 σ^2 的无偏估计。

利用逐步回归算法,在保留相同预报因子个数的条件下,得到伪造因变量 y 的回归方程,求得复相关系数 R 的抽样值。重复蒙特卡洛模拟,得到 R 的分布。

对给定的显著水平 α,若

$$P\{R>R^*\}>\alpha$$

则所得回归预报和随机预报相比差异并不显著,在实际预报中无法应用,可拒绝接受回归方程。

无疑,这种和随机预报相比的模拟方法,在时间序列分析、信号检测、滤波效果检查中都是可用的。

(2) 放大器线路分析

对一个复杂的系统,要建立一个真实的可靠性分析模型,是十分困难的。用蒙特卡洛方法模拟这类系统,可以大大节省工程费用和设计时间。下面的放大器线路分析模型,可作为一个简单的说明性例子[10]。

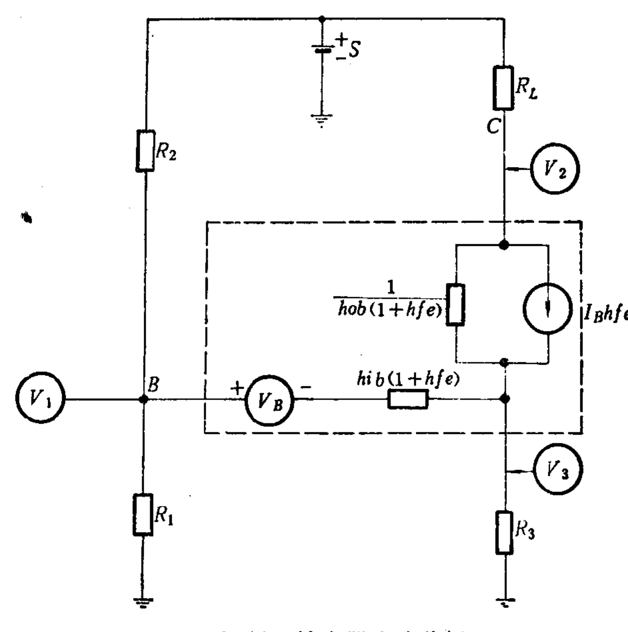


图7.11 放大器线路分析

在图7.11的放大器线路中,假定线路中所有参量都是正态分布的,具有已知的均值和标准差,其数值由下表给出。

参　　量	均　　值	标　准　差
R_1	1000 Ω	16.7 Ω
R_2	6000 Ω	100 Ω
R_3	500 Ω	8.33 Ω
$R_L=R_4$	2000 Ω	33.3 Ω
$\frac{1}{hob(1+hfe)}=R_5$	43000 Ω	2150 Ω
$hib(1+hfe)=R_6$	487.5 Ω	24.4 Ω
$hfe=h$	42	4.2
V	0.58 V	0 V
S	20 V	0.25 V

假若参量 V_1 可以确定系统的可靠性。设 1.30 伏 $<V_1<$ 1.43 伏是线路正常工作的临界区域,要求计算线路的可靠性。

分析放大器线路图7.11,得到线性方程组

$$\left.\begin{aligned}
&\left(\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_6}\right)V_1-\frac{1}{R_6}V_3=\frac{S}{R_2}+\frac{V}{R_6}\\
&\frac{h}{R_6}V_1+\left(\frac{1}{R_4}+\frac{1}{R_5}\right)V_2-\left(\frac{1}{R_5}+\frac{h}{R_6}\right)V_3=\frac{S}{R_4}+\frac{Vh}{R_6}\\
&\frac{1+h}{R_6}V_1-\frac{1}{R_5}V_2+\left(\frac{1}{R_3}+\frac{1+h}{R_6}+\frac{1}{R_5}\right)V_3=-\frac{V(1+h)}{R_6}
\end{aligned}\right\} \tag{7.8.2}$$

求解随机线代数方程(7.8.2)，得 V_1:

$$V_1=\frac{\begin{vmatrix}
\dfrac{SR_6+VR_2}{R_2R_6} & 0 & -\dfrac{1}{R_6}\\
\dfrac{SR_6+VhR_4}{R_4R_6} & \dfrac{R_4+R_5}{R_4R_5} & -\dfrac{hR_5+R_6}{R_5R_6}\\
-\dfrac{V(1+h)}{R_6} & -\dfrac{1}{R_5} & \dfrac{R_3R_5(1+h)+(R_3+R_5)R_6}{R_3R_5R_6}
\end{vmatrix}}{\begin{vmatrix}
\dfrac{R_1R_2+R_1R_6+R_2R_6}{R_1R_2R_6} & 0 & -\dfrac{1}{R_6}\\
\dfrac{h}{R_6} & \dfrac{R_4+R_5}{R_4R_5} & -\dfrac{hR_5+R_6}{R_5R_6}\\
\dfrac{1+h}{R_6} & -\dfrac{1}{R_5} & \dfrac{R_3R_5(1+h)+(R_3+R_5)R_6}{R_3R_5R_6}
\end{vmatrix}}$$

在计算机上，用§7.3给出的抽样方法，产生正态随机变量，可给出随机变量 V_1 的模拟值，算出放大器线路的可靠性。

在计算机上[10]，以置信水平75%，算得可靠性为90%。

(3) 控制棒吸收率的统计模拟

质点随机游动的统计模拟问题，是蒙特卡洛方法应用的重要领域之一[1,14]，可用来解决中子物理和核反应堆物理中提出的不少数值问题。作为例子，讨论热中子在控制棒内随机游动的模拟问题，计算控制棒的吸收率。

控制棒是核反应堆的一个重要组成部分，外形是一个长为 L、半径为 R 的圆柱体，横截面如图7.12所示。

控制棒中间是一个半径为 r 的碳化硼圆柱体，以概率1吸收进入它内部的热中子。在碳化硼外面，是一层厚为 l_1 的铝合金包壳。包壳外面有厚为 l_2 的冷却水层。水层外面，又用铝合金作包壳，厚为 l_3。这里，$l_1+l_2+l_3+r=R$。

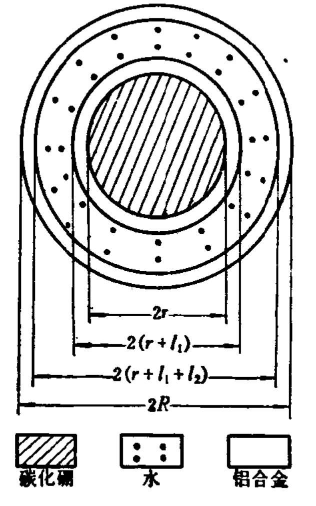


图7.12 控制棒横截面

由于 $L\gg R$，为简化模型，视控制棒为无限长。这时，入射控制棒的中子与入射点的位置无关。

热中子进入控制棒的入射角 $(\theta,\ \varphi)$ 是一对随机变量，密度函数

$$f(\theta,\ \varphi)=\begin{cases}\dfrac{1}{C}\sin^2\theta\sin\varphi(1+\sqrt{s}\sin\theta\sin\varphi), & \text{当 } 0<\theta,\ \ \varphi<\dfrac{\pi}{2}\text{ 时}\\ 0, & \text{其它}\end{cases}$$

在直角坐标系中，热中子的入射方向

$$\begin{cases}\Delta x=\sin\theta\cos\varphi\\ \Delta y=\sin\theta\sin\varphi\\ \Delta z=\cos\theta\end{cases}$$

抽样方法由例 9、例 11 给出。

热中子在介质内随机游动时，和介质内的质点发生碰撞，或以概率 $P_s=\Sigma_s/\Sigma_t$ 散射，或以概率 $P_a=\Sigma_a/\Sigma_t$ 吸收。这里，Σ_s，Σ_a，$\Sigma_t=\Sigma_s+\Sigma_a$，分别表示碰撞点处介质的散射截面、吸收截面和总截面，它们都随介质的不同而异。模拟时，产生随机数 r，视其是否小于 P_s，确定碰撞的类型(§7.3.1)。

热中子在水中发生散射时，散射角 ω 以

$$f(\omega)=\begin{cases}\dfrac{1}{\pi}(1+3\mu_0\cos\omega), & 0<\omega<\pi\\ 0, & \text{其它}\end{cases}$$

为密度函数，方位角 φ 均匀分布在 $[0,\ 2\pi]$ 上。$\cos\omega$ 和 $\cos\varphi$ 的抽样方法，已由例 5、例 6 给出。

若散射前中子的游动方向为 $(\Delta x,\ \Delta y,\ \Delta z)$，散射后新的游动方向为 $(\Delta x',\ \Delta y',\ \Delta z')$，则有

$$\begin{cases}\Delta z'=\Delta z\cos\omega+\sin\omega\cos\varphi\sqrt{1-\Delta z^2}\\ \Delta y'=\dfrac{1}{1-\Delta z^2}\left[\Delta y(\cos\omega-\Delta z\Delta z')+\Delta x\sin\omega\sin\varphi\sqrt{1-\Delta z^2}\right]\\ \Delta x'=\dfrac{1}{1-\Delta z^2}\left[\Delta x(\cos\omega-\Delta z\Delta z')-\Delta y\sin\omega\sin\varphi\sqrt{1-\Delta z^2}\right]\end{cases}$$

在铝中，发生散射的中子是各向同性的。取

$$\lambda=\frac{\sqrt{3}}{2\sqrt[3]{2}}$$

给出速度较快的一种抽样方法，如图 7.13 所示。

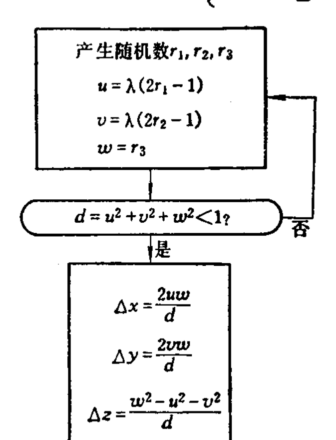


图 7.13　抽样框图

热中子在介质内发生散射后，继续游动。游动长度 l 服从以 Σ_t 为参数的指数分布(参见例 2)，即中子游动的平均自由程为 $\dfrac{1}{\Sigma_t}$。

由于我们模拟的中子要在不同的介质内游动，Σ_t 是变化的，l 的抽样过程较为复杂。沿中子的散射方向，依次算出在各个不同介质里需要游动的长度，分别记为 $l_1,\ l_2,\ \cdots,\ l_k$，相应各介质的总截面 Σ_t 分别记为 $\Sigma_1,\ \Sigma_2,\ \cdots,\ \Sigma_k$。游动长度 l 的抽样方法为：

产生随机数 r，计算标准指数分布的随机变量 $\eta=-\ln r$。找出满足不等式

$$\Sigma_1 l_1+\cdots+\Sigma_{i-1}l_{i-1}<-\ln r<\Sigma_1 l_1+\cdots+\Sigma_{i-1}l_{i-1}+\Sigma_i l_i$$

的标号值 i，计算游动自由程

$$l=l_1+\cdots+l_{i-1}+\frac{-\ln r-(\Sigma_1 l_1+\cdots+\Sigma_{i-1}l_{i-1})}{\Sigma_i}$$

如果中子的散射点为 $\boldsymbol{r}_0$，散射方向为 $\boldsymbol{\omega}$，则游动中子新的碰撞点为

$$\boldsymbol{r}=\boldsymbol{r}_0+l\boldsymbol{\omega}$$

当

$$-\ln r>\Sigma_1 l_1+\cdots+\Sigma_k l_k$$

时，质点飞出游动介质的区域。

根据上面给出的概型和抽样方法，可跟踪热中子至吸收或飞出控制棒为止。研究控制棒的吸收率，即计算热中子被水、铝合金和碳化硼的总吸收率。无疑，这里给出的是一种直接模拟方法。为了提高模拟结果的收敛速度，根据§7.7给出的加速收敛原理和降低模拟方差的一些技巧，可采用新的一组模拟方法。

(i) 分层抽样

把中子的入射区域(如图7.14)分成三部分。按各层的贡献和方差的大小，用分层抽样的方法，在各层中抽样不同的中子数。

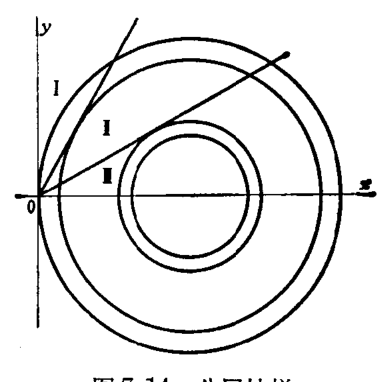


图7.14 分层抽样

(ii) 统计加权抽样

代替质点吸收、散射的抽样方法，用期望评分来表示。设一个质量为 W_n 的质点和介质中的粒子发生第 $(n+1)$ 次碰撞，用期望评分、统计加权的方法，则有 $P_a W_n$ 的质量被控制棒吸收，而中子以 $W_{n+1}=P_s W_n$ 的质量继续游动。

(iii) 统计估计抽样

在进行自由程 l 的抽样时，用改变游动中子的质量，代替质点飞出控制棒的统计模拟。在第 n 次碰撞后，有

$$W_n \exp[-(\Sigma_1 l_1+\Sigma_2 l_2+\cdots+\Sigma_k l_k)]$$

的质量飞出控制棒。这时，继续游动的中子的质量

$$W'_n=W_n\{1-\exp[-(\Sigma_1 l_1+\Sigma_2 l_2+\cdots+\Sigma_k l_k)]\}$$

代替随机数 r，取

$$r^*=1-r\{1-\exp[-(\Sigma_1 l_1+\Sigma_2 l_2+\cdots+\Sigma_k l_k)]\}$$

计算 $-\ln r^*$，抽样游动自由程 l。

最后，把用高维积分计算得到的近似数值结果和进行25000次直接模拟得到的统计结果，列表比较如下：

数值 / 算法 \ 类别	热中子进入各区的频率			进入各区热中子的吸收率			控制棒的吸收率
	I	II	III	I	II	III	
积分结果	0.061165	0.263035	0.6758	0	0.4192	0.6387	0.512
模拟结果	0.05992	0.26392	0.67616	0.0527	0.381	0.6146	0.5188

参 考 资 料

[1] Н. П. 布斯连科, Ю. А. 施廖盖尔著, 杜淑敏等译,《统计试验法(蒙特卡洛法)》, 科学技术出版社, 1964.

[2] Б. В. 格涅坚科著, 丁寿田译,《概率论教程》, 人民教育出版社, 1960。

[3] Franklin J. N., "Numerical simulation of stationary and nonstationary Gaussian random processes", *SIAM. Rev.*, **7**(1), 1965, pp 68～80.

[4] Halton J. H., "A retrospective and prospective survey of the monte carlo methods", *SIAM. Rev.*, **12**(1), 1970, pp 1～63.

[5] Hull T. E., Dobell A. R., "Random number generator" *SIAM. Rev.*, **4**(3), 1962, pp 230～254.

[6] Lin S. C., "Statistical analysis and stochastic simulation of ground motion data", *The Bell Systems Technical J.*, **47**(10), 1968, pp 2273～2298.

[7] Lund I. A., "A Monte Carlo method for testing the statistical significance of a regression equation", J. of Appl. Meteor., **9**(3), 1970, pp 330～332.

[8] Meyer H. A., ed., "Symposium on Monte Carlo Methods", John Wiley, 1956.

[9] Muller M. E., "A comparrison of methods for generating normal deviates on digital computers", *JACM.*, **6**(3), 1959, pp 376～383.

[10] Myers P. J., "Monte Carlo: Reliability tool for design engineers", Proceeding of Ninth National Symposium on Reliability & Quality Control, 1963, pp 487～492.

[11] Naylor T. H., ed., "Computer Simulation Technique", John. Wiley, 1966.

[12] Rand Corporation, "A Million Random Digits with 100,000 Normal Deviates", The Free Press, 1955.

[13] Tocher K. D., "The Art of Simulation", The English Universities Press, 1963.

[14] Шрейдер Ю. А. и др., "Метод Статистигеских Испытаний" (Метод Монте-Карло), Физматгиз, 1962.

[15] Полляк Ю. Г., "Вераятностное Моделирование На Электронных Вычислительных Машинах", Советское Радио, 1971.

第八章　线性代数方程组的数值解法

线性代数方程组的数值求解是工程实践中经常遇到的问题。据不完全统计，工程实践中提出的计算问题，有一半以上包括求解线性代数方程组。例如，结构应力分析问题、电力传输网分析问题、大地测量问题、数据拟合问题、各种晶体管电路分析问题以及非线性方程组与微分方程数值解问题等等。因而，了解在计算机上用什么方法求解线性代数方程组，对于使用计算机来说是十分必要的。本章介绍一些目前在计算机上经常使用的、简单有效的方法。

通常把线性代数方程组的数值解法分为直接法与迭代法两类。所谓直接法，就是经有限次数的运算即可求得(如果没有舍入误差)方程组精确解的方法。迭代法则与之相反，它将求解方程组的问题化为构造一个无限序列，其极限就是方程组的解答，因而在有限步内是得不到精确解的。直接法与迭代法各有优缺点，前者由于受到计算机存储容量的限制，一般来说仅适宜于系数矩阵阶数不是太高的问题，其工作量较小，精确度较高，但程序较复杂。后者主要用于某些高阶问题，特别是在椭圆型偏微分方程边值问题的数值求解中有着广泛的应用。一般来说，其程序较为简单，但工作量有时较大。实际计算时，应根据问题的特点和要求以及所使用的计算机的情况，来决定方法的取舍。

我们在§8.1中介绍一些常用的直接解法，并在本章最后给出相应的算法语言程序，这些程序都在计算机上进行过试算，感兴趣的读者可以直接引用它们。§8.2是关于迭代法某些基本概念的介绍。由于迭代法主要用于椭圆型偏微分方程边值问题的数值解，本书第十三章中将详细讨论常用的迭代法的应用问题，这里仅从矩阵计算的角度作一些简单的讨论。最后，在§8.3中介绍一些解线性矛盾方程组的常用方法。

§8.1　解线性代数方程组的直接法

本节介绍计算机上常用的解线性代数方程组的直接法。一个 n 阶线性代数方程组可以表示为：

$$\begin{cases} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n=b_1 \\ a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n=b_2 \\ \cdots\cdots \\ a_{n1}x_1+a_{n2}x_2+\cdots+a_{nn}x_n=b_n \end{cases} \tag{8.1.1}$$

采用矩阵和向量符号，这个方程组就可写为：

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \tag{8.1.2}$$

其中 $\boldsymbol{A}=[a_{ij}]$ 是由方程组(8.1.1)的系数 a_{ij} 所构成的 $n\times n$ 矩阵，通常称为系数矩阵；$\boldsymbol{b}=[b_i]$ 为方程组(8.1.1)的右端项所构成的 n 维向量，通常称之为右端向量或自由项；$\boldsymbol{x}=[x_i]$ 为要求的解答 x_i 所构成的 n 维向量。

大家知道，只要矩阵 A 的行列式 $\det\boldsymbol{A}\neq 0$，方程组(8.1.1)就有唯一解，其表达式为：

$$x_i=\det\boldsymbol{A}_i/\det\boldsymbol{A}\quad(i=1,\ 2,\ \cdots,\ n) \tag{8.1.3}$$

其中 $\boldsymbol{A}_i$ 为用右端向量 $\boldsymbol{b}$ 替换系数矩阵 $\boldsymbol{A}$ 的第 i 列而得的矩阵。这一公式称之为克莱姆法则。显然，按克莱姆法则求解方程组(8.1.1)，需要计算 $n+1$ 个 n 阶行列式。每个 n 阶行列式按直接展开办法来算需 $(n-1)\times n!$ 次乘法和 $n!$ 次加法运算。当 $n=30$ 时，共约需完成 10^{33} 次乘法和加法运算，这是一个十分惊人的数字，即使在一台每秒作一亿次运算的计算机上完成这一计算也是不可能的。所以，尽管这种办法也是一种直接法，并且理论上可行，但实际上是无法进行求解的。即使采用其它办法来计算行列式，按克莱姆法则求解的工作量也比通常的直接法大得多。因而，克莱姆法则对于数值计算来说是没有什么用处的，仅在一些特殊场合才有用。

现在计算机上常用的直接解法大多数是以系数矩阵的三角形化为基础的。即是说，先对方程组进行变换，使其化为等价的(即具有相同解答的)三角形方程组。由于三角形方程组的求解十分容易，原来方程组的求解问题即告解决。本节中只讨论这一类型的方法。为讨论方便起见，我们首先叙述三角形方程组的解法，然后叙述各种将原方程组化为等价三角形方程组的方法，并比较其优劣。由于计算机的字长是有限的，每次运算之后还要对结果进行舍入，所以，虽然理论上直接法在有限步内可以得到精确解，但实际在计算机上得到的只是近似解。这样就产生了误差分析的问题。本节的最后一段将简单地讨论一下这个问题。

8.1.1 三角形方程组的解法

所谓三角形方程组是指下面两种形式的方程组：

$$
\begin{cases}
l_{11}x_1 & =b_1 \\
l_{21}x_1+l_{22}x_2 & =b_2 \\
l_{31}x_1+l_{32}x_2+l_{33}x_3 & =b_3 \\
\cdots\cdots & \\
l_{n1}x_1+l_{n2}x_2+l_{n3}x_3+\cdots+l_{nn}x_n & =b_n
\end{cases}
\tag{8.1.4}
$$

$$
\begin{cases}
u_{11}x_1+u_{12}x_2+u_{13}x_3+\cdots+u_{1n}x_n=d_1 \\
\qquad u_{22}x_2+u_{23}x_3+\cdots+u_{2n}x_n=d_2 \\
\qquad\qquad \cdots\cdots \\
\qquad\qquad u_{n-1,n-1}x_{n-1}+u_{n-1,n}x_n=d_{n-1} \\
\qquad\qquad\qquad u_{nn}x_n=d_n
\end{cases}
\tag{8.1.5}
$$

方程组(8.1.4)叫作下三角形方程组，(8.1.5)叫作上三角形方程组。若用矩阵符号即可分别写为：

$$\boldsymbol{L}\boldsymbol{x}=\boldsymbol{b} \quad \text{或} \quad \boldsymbol{U}\boldsymbol{x}=\boldsymbol{d}$$

其中 $\boldsymbol{L}=[l_{ij}]$ 为方程组(8.1.4)的系数所构成的下三角形矩阵，其元素满足关系：

$$l_{ij}=0 \quad (i<j)$$

$\boldsymbol{U}$ 为方程组(8.1.5)的系数所构成的上三角形矩阵，其元素满足关系式：

$$u_{ij}=0 \quad (i>j)$$

三角形方程组的求解是很简单的。例如，对于方程组(8.1.4)，我们可以从其第一个方程定出 $x_1=b_1/l_{11}$，然后将它代入第二个方程，便可求得 $x_2=(b_2-l_{21}x_1)/l_{22}$。把求得的 x_1，x_2 代入第三个方程又可定出 x_3，如此逐个方程地向前递推下去，n 步之后即可求得全部解答。这一过程有时也叫作前推过程。显然，其计算公式可以归结为：

$$\begin{cases} x_1 = b_1/l_{11} \\ x_i = (b_i - l_{i1}x_1 - l_{i2}x_2 - \cdots - l_{i,i-1}x_{i-1})/l_{ii} \quad (i=2,\ 3,\ \cdots,\ n) \end{cases} \tag{8.1.6}$$

求出x_i需要作$i-1$次乘法和加减法以及一次除法，故总共需完成$1+2+3+\cdots+n-1\approx n^2/2$次乘法和加减法以及$n$次除法。

完全类似，对于方程组(8.1.5)，我们可以先从第n个方程定出$x_n=d_n/u_{nn}$，然后将它代入第$n-1$个方程定出$x_{n-1}=(d_{n-1}-u_{n-1,n}x_n)/u_{n-1,n-1}$。如此逐个方程地回代下去，最终即可求出全部解答。这个计算所需过程通常叫作回代过程。显然，其计算公式可以归结为：

$$\begin{cases} x_n = d_n/u_{n,n} \\ x_i = (d_i - u_{i,i+1}x_{i+1} - u_{i,i+2}x_{i+2} - \cdots - u_{i,n}x_n)/u_{i,i} \quad (i=n-1,\ n-2,\ \cdots,\ 1) \end{cases} \tag{8.1.7}$$

回代过程所需完成的计算量也是$n^2/2$次乘法和加减法以及n次除法。

从(8.1.6)和(8.1.7)可以看出，求解三角形方程组是很简单的，只要把方程组化成了等价的三角形方程组，求解即很容易完成。下面我们就来叙述各种将原方程组化为等价三角形方程组的方法，并比较其优劣。

8.1.2 高斯消去法

(一)高斯消去法的基本步骤

高斯消去法(以后简称为消去法)是大多数读者早已熟悉的方法，它的提出已有相当长的时间了，故是一个古老的方法。然而，近年来在计算机上求解线性代数方程组的实践表明，它仍是直接法中最常用的一种方法，也是最有效的方法之一。其基本思想是用逐次消去一个未知数的办法把原来的方程组化为等价的(即具有相同解答的)三角形方程组，这样，解答就很容易求得。为了说明简便，我们以下面的三阶方程组为例：

$$\begin{cases} 4x_1 - 9x_2 + 2x_3 = 5 \\ 2x_1 - 4x_2 + 6x_3 = 3 \\ x_1 - x_2 + 3x_3 = 4 \end{cases} \quad \text{或} \quad \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$$

首先我们来消去第二、三两个方程中的未知数x_1。为此，只需将第一个方程分别乘以$-2/4$及$-1/4$，加至第二及第三个方程上去。于是方程组就变为：

$$\begin{cases} 4x_1 - 9x_2 + 2x_3 = 5 \\ 0.5x_2 + 5x_3 = 0.5 \\ 1.25x_2 + 2.5x_3 = 2.75 \end{cases} \quad \text{或} \quad \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 1.25 & 2.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 0.5 \\ 2.75 \end{bmatrix}$$

显然，这一方程组的解答与原方程组是相同的。

完全类似，将所得方程组的第二个方程式乘以$-1.25/0.5$，并加至第三个方程式上，即可消去其未知数x_2。于是得出如下等价三角形方程组：

$$\begin{cases} 4x_1 - 9x_2 + 2x_3 = 5 \\ 0.5x_2 + 5x_3 = 0.5 \\ -10x_3 = 1.5 \end{cases} \quad \text{或} \quad \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 0 & -10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 0.5 \\ 1.5 \end{bmatrix}$$

这个等价三角形方程组很容易由前述的回代过程(8.1.7)求解，即从其第三个方程求得$x_3=-0.15$。将它代入第二个方程即求得$x_2=2.5$。最后从第一个方程式求得$x_1=6.95$。这样就完成了消去法的整个求解过程。

对于一般的n阶线性方程组，消去法的计算步骤也是完全类似的。假定把要求解的n阶

线性代数方程组(8.1.1)改写为如下形式:

$$\begin{cases} a_{11}^{(1)}x_1+a_{12}^{(1)}x_2+a_{13}^{(1)}x_3+\cdots+a_{1n}^{(1)}x_n=b_1^{(1)} \\ a_{21}^{(1)}x_1+a_{22}^{(1)}x_2+a_{23}^{(1)}x_3+\cdots+a_{2n}^{(1)}x_n=b_2^{(1)} \\ \cdots\cdots \\ a_{n1}^{(1)}x_1+a_{n2}^{(1)}x_2+a_{n3}^{(1)}x_3+\cdots+a_{nn}^{(1)}x_n=b_n^{(1)} \end{cases} \tag{8.1.8}$$

或用矩阵符号记为:

$$\boldsymbol{A}^{(1)}\boldsymbol{x}=\boldsymbol{b}^{(1)}$$

其中 $\boldsymbol{A}^{(1)}$ 为 $n\times n$ 方阵; $\boldsymbol{b}^{(1)}$ 为 $n\times 1$ 向量; 它们分别为:

$$\boldsymbol{A}^{(1)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ \cdots\cdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} \end{bmatrix}, \quad \boldsymbol{b}^{(1)}=\begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{bmatrix}$$

假设 $a_{11}^{(1)}\neq 0$, 分别从原方程组的第二个方程减去第一个方程乘以 $a_{21}^{(1)}/a_{11}^{(1)}$, 第三个方程减去第一个方程乘以 $a_{31}^{(1)}/a_{11}^{(1)}$, 如此等等, 即可消去后面 $n-1$ 个方程中的未知数 x_1。这时, 方程组(8.1.8)就变为如下等价方程组:

$$\begin{cases} a_{11}^{(1)}x_1+a_{12}^{(1)}x_2+a_{13}^{(1)}x_3+\cdots+a_{1n}^{(1)}x_n=b_1^{(1)} \\ \qquad a_{22}^{(2)}x_2+a_{23}^{(2)}x_3+\cdots+a_{2n}^{(2)}x_n=b_2^{(2)} \\ \qquad a_{32}^{(2)}x_2+a_{33}^{(2)}x_3+\cdots+a_{3n}^{(2)}x_n=b_3^{(2)} \\ \qquad \cdots\cdots \\ \qquad a_{n2}^{(2)}x_2+a_{n3}^{(2)}x_3+\cdots+a_{nn}^{(2)}x_n=b_n^{(2)} \end{cases}$$

或者用矩阵符号记为:

$$\boldsymbol{A}^{(2)}\boldsymbol{x}=\boldsymbol{b}^{(2)}$$

其中

$$\boldsymbol{A}^{(2)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \cdots\cdots \\ 0 & a_{n2}^{(2)} & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}, \quad \boldsymbol{b}^{(2)}=\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}$$

若令 $m_{i1}=a_{i1}^{(1)}/a_{11}^{(1)}$, 则系数 $a_{ij}^{(2)}$ 和 $b_i^{(2)}$ 的计算公式应为:

$$a_{ij}^{(2)}=a_{ij}^{(1)}-m_{i1}\cdot a_{1j}^{(1)}, \quad b_i^{(2)}=b_i^{(1)}-m_{i1}\cdot b_1^{(1)} \quad (i,\ j=2,\ 3,\ \cdots,\ n)$$

显然, 这一步消去需要完成 $n\times(n-1)$ 次乘法和加减法运算。

类似地, 若 $a_{22}^{(2)}\neq 0$, 分别从上述等价方程组的第三个方程减去第二个方程乘以 $a_{32}^{(2)}/a_{22}^{(2)}$, 第四个方程减去第二个方程乘以 $a_{42}^{(2)}/a_{22}^{(2)}$, 如此等等, 即可进一步消去后面 $n-2$ 个方程中的未知数 x_2, 而将方程组(8.1.8)变为如下等价形式:

$$\begin{cases} a_{11}^{(1)}x_1+a_{12}^{(1)}x_2+a_{13}^{(1)}x_3+\cdots+a_{1n}^{(1)}x_n=b_1^{(1)} \\ \qquad a_{22}^{(2)}x_2+a_{23}^{(2)}x_3+\cdots+a_{2n}^{(2)}x_n=b_2^{(2)} \\ \qquad\qquad a_{33}^{(3)}x_3+\cdots+a_{3n}^{(3)}x_n=b_3^{(3)} \\ \qquad\qquad a_{43}^{(3)}x_3+\cdots+a_{4n}^{(3)}x_n=b_4^{(3)} \\ \qquad\qquad \cdots\cdots \\ \qquad\qquad a_{n3}^{(3)}x_3+\cdots+a_{nn}^{(3)}x_n=b_n^{(3)} \end{cases}$$

或用矩阵符号记为:

$$A^{(3)}x=b^{(3)}$$

其中

$$A^{(3)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ \vdots & 0 & a_{43}^{(3)} & \cdots & a_{4n}^{(3)} \\ \vdots & \vdots & \cdots\cdots & & \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} \end{bmatrix},\quad b^{(3)}=\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ b_4^{(3)} \\ \vdots \\ b_n^{(3)} \end{bmatrix}$$

若令 $m_{i2}=a_{i2}^{(2)}/a_{22}^{(2)}$,则系数 $a_{ij}^{(3)}$ 和 $b_i^{(3)}$ 的计算公式应为:

$$a_{ij}^{(3)}=a_{ij}^{(2)}-m_{i2}a_{2j}^{(2)},\quad b_i^{(3)}=b_i^{(2)}-m_{i2}b_2^{(2)}\quad (i,\ j=3,\ 4,\ \cdots,\ n)$$

容易验证,这一步消去需完成$(n-1)\times(n-2)$次乘法和加减法运算。

如果 $a_{33}^{(3)}\neq 0$, 上述的消去步骤还可以进行下去。如此继续之,重复上述步骤$(n-1)$次以后,我们即可得到如下等价三角形方程组:

$$\begin{cases} a_{11}^{(1)}x_1+a_{12}^{(1)}x_2+a_{13}^{(1)}x_3+\cdots+a_{1n}^{(1)}x_n=b_1^{(1)} \\ \qquad a_{22}^{(2)}x_2+a_{23}^{(2)}x_3+\cdots+a_{2n}^{(2)}x_n=b_2^{(2)} \\ \qquad\qquad a_{33}^{(3)}x_3+\cdots+a_{3n}^{(3)}x_n=b_3^{(3)} \\ \qquad\qquad\qquad \ddots \qquad \vdots \qquad \vdots \\ \qquad\qquad\qquad\qquad a_{nn}^{(n)}x_n=b_n^{(n)} \end{cases} \tag{8.1.9}$$

或者用矩阵符号记为:

$$A^{(n)}x=b^{(n)}$$

其中 $A^{(n)}$ 为如下的上三角形矩阵, $b^{(n)}$ 为 $n\times 1$ 向量:

$$A^{(n)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ & \mathbf{0} & & \ddots & \vdots \\ & & & & a_{nn}^{(n)} \end{bmatrix},\quad b^{(n)}=\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_n^{(n)} \end{bmatrix} \tag{8.1.10}$$

三角形方程组 $A^{(n)}x=b^{(n)}$ 很容易用前述的回代过程(8.1.7)求解,这就完成了消去法求解 n 阶线性代数方程组的过程。从原来方程组(8.1.8)得出等价三角形方程组(8.1.9)的过程通常称之为消去过程。采用前面的记号,我们可将消去过程的计算公式归结为对于 $k=1,\ 2,\ \cdots,\ n-1$, 递推地计算如下各量:

$$\begin{cases} a_{ij}^{(k+1)}=a_{ij}^{(k)} \\ b_i^{(k+1)}=b_i^{(k)} \end{cases}\ (i\leqslant k,\ j\leqslant n) \\ \begin{cases} a_{ij}^{(k+1)}=a_{ij}^{(k)}-\left(\dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right)\cdot a_{kj}^{(k)} \\ b_i^{(k+1)}=b_i^{(k)}-\left(\dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right)\cdot b_k^{(k)} \end{cases}\ (k+1\leqslant i\leqslant n,\ k+1\leqslant j\leqslant n) \\ a_{ij}^{(k+1)}=0\quad (1\leqslant j\leqslant k,\ j+1\leqslant i\leqslant n) \tag{8.1.11}$$

由于消去过程的第 k 步所需完成的乘法和加减法量为: $(n-k+1)\times(n-k)$, 所以,消去过程(8.1.11)共需完成 $\sum\limits_{k=1}^{n-1}(n-k)\times(n-k+1)\sim\dfrac{1}{3}n^3$ 次乘法和加减法运算。从前面知道,

解三角形方程组(8.1.9)所需的乘法与加减法量约为$n^2/2$，即是说，用消去法求解n阶线性方程组所需的乘法与加减法量大约为：$\frac{1}{3}n^3+O(n^2)$，其中主要的运算量$\left(\frac{1}{3}n^3\right)$花费在消去过程上面。

(二)系数矩阵的三角形分解

现在我们从矩阵运算的角度来考察一下前面的消去过程。按照矩阵乘法规则，读者很容易验证，消去过程的第一步，系数矩阵和自由项的变化相当于进行如下矩阵乘法：

$$\boldsymbol{A}^{(2)}=\boldsymbol{M}_1\cdot\boldsymbol{A}^{(1)},\quad \boldsymbol{b}^{(2)}=\boldsymbol{M}_1\cdot\boldsymbol{b}^{(1)}$$

其中

$$\boldsymbol{M}_1=\begin{bmatrix}1 & & & \\ -m_{21} & 1 & & \mathbf{0} \\ -m_{31} & & 1 & \\ \vdots & \mathbf{0} & & \ddots \\ -m_{n1} & & & & 1\end{bmatrix}$$

$$m_{i1}=a_{i1}^{(1)}/a_{11}^{(1)}\quad (i=2,\ 3,\ \cdots,\ n)$$

类似地，消去过程第二步，系数矩阵和自由项的变化可以表为：

$$\boldsymbol{A}^{(3)}=\boldsymbol{M}_2\cdot\boldsymbol{A}^{(2)},\quad \boldsymbol{b}^{(3)}=\boldsymbol{M}_2\cdot\boldsymbol{b}^{(2)}$$

其中

$$\boldsymbol{M}_2=\begin{bmatrix}1 & & & & \\ 0 & 1 & & \mathbf{0} & \\ 0 & -m_{32} & 1 & & \\ 0 & -m_{42} & & 1 & \\ \vdots & \vdots & \mathbf{0} & & \ddots \\ 0 & -m_{n2} & & & & 1\end{bmatrix}$$

$$m_{i2}=a_{i2}^{(2)}/a_{22}^{(2)}\quad (i=3,\ 4,\ \cdots,\ n)$$

即是说，每消去一步相当于在方程组的系数矩阵与自由项上分别左乘以相应的矩阵$\boldsymbol{M}_i$。这样，$n-1$步之后即得到与(8.1.9)相应的上三角形系数矩阵$\boldsymbol{A}^{(n)}$和自由项$\boldsymbol{b}^{(n)}$。所以，我们有：

$$\begin{cases}\boldsymbol{A}^{(n)}=\boldsymbol{M}_{n-1}\cdot\boldsymbol{M}_{n-2}\cdot\cdots\cdot\boldsymbol{M}_2\cdot\boldsymbol{M}_1\cdot\boldsymbol{A}^{(1)} \\ \boldsymbol{b}^{(n)}=\boldsymbol{M}_{n-1}\cdot\boldsymbol{M}_{n-2}\cdot\cdots\cdot\boldsymbol{M}_2\cdot\boldsymbol{M}_1\cdot\boldsymbol{b}^{(1)}\end{cases}\tag{8.1.12}$$

其中$\boldsymbol{M}_i$为第i列对角线以下元素为非零，而其它元素与单位矩阵相同的下三角形矩阵，通常称之为初等变换矩阵。它的第i列的非零元素为：$-m_{ji}=-a_{ji}^{(i)}/a_{ii}^{(i)}$，$(j=i+1,\ i+2,\ \cdots,\ n)$，亦即：

$$\boldsymbol{M}_i=\begin{bmatrix}1 & & & & & \\ & 1 & & & \mathbf{0} & \\ & & \ddots\ 1 & & & \\ & & -m_{i+1,i} & 1 & & \\ \mathbf{0} & & -m_{i+2,i} & & 1 & \\ & & \vdots & \mathbf{0} & & \ddots \\ & & -m_{n,i} & & & & 1\end{bmatrix}\tag{8.1.13}$$

因此，消去过程实质上就是用一系列初等变换矩阵(8.1.13)来左乘原来方程组的系数矩阵，而将其化为上三角形矩阵的过程，也叫作用初等变换矩阵的三角化过程。这样就把求解一般的n阶线性方程组变为求解等价的三角形方程组。

此外，按照逆矩阵的定义，读者很容易验证：

$$M_i^{-1}=\begin{bmatrix}1 & & & & & & \\ & 1 & & & & & \mathbf{0} \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & m_{i+1,i} & 1 & & \\ & \mathbf{0} & & m_{i+2,i} & & \ddots & \\ & & & \vdots & & & 1 \\ & & & m_{n,i} & \mathbf{0} & & \end{bmatrix} \tag{8.1.14}$$

所以，从(8.1.12)我们有：

$$A^{(1)}=M_1^{-1}\cdot M_2^{-1}\cdot\cdots\cdot M_{n-2}^{-1}\cdot M_{n-1}^{-1}\cdot A^{(n)}$$

若令 $U=A^{(n)}$, $L=M_1^{-1}\cdot M_2^{-1}\cdot\cdots\cdot M_{n-1}^{-1}$, 也很容易用矩阵乘法规则验证：

$$L=M_1^{-1}\cdot M_2^{-1}\cdots M_{n-1}^{-1}=\begin{bmatrix}1 & & & & \\ m_{21} & 1 & & \mathbf{0} & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & & \ddots & 1 \\ m_{n1} & m_{n2} & \cdots & m_{n,n-1} & 1\end{bmatrix}$$

即是说，我们有

$$A^{(1)}=L\cdot U=\begin{bmatrix}1 & & & \\ m_{21} & 1 & \mathbf{0} & \\ m_{31} & m_{32} & 1 & \\ \vdots & \vdots & & \ddots \\ m_{n1} & m_{n2} & & 1\end{bmatrix}\cdot\begin{bmatrix}a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ \mathbf{0} & & & a_{nn}^{(n)}\end{bmatrix}$$

这样就得到了系数矩阵 $A^{(1)}$ 的三角形分解式。

从消去法的计算过程得知，只要逐次的主对角线元素 $a_{kk}^{(k)}\neq 0$, 消去过程总是可以进行下去的，也就是说，总可得到上述的分解式。此外，由于左乘以矩阵 M_i 后矩阵的行列式值不变，并且矩阵 $A^{(1)}$ 的每一行在消去过程中只和其前面各行进行运算，所以，还可看出，系数矩阵 $A^{(1)}$ 的逐次左上角主子式 $\det A_k^{(1)}$ 应满足下列关系：

$$\det A_k^{(1)}=\det A_k^{(k)}=a_{11}^{(1)}\cdot a_{22}^{(2)}\cdots a_{kk}^{(k)}\ (k=1,\ 2,\ \cdots,\ n)$$

只要 $\det A_k^{(1)}\neq 0(k=1,\ 2,\ \cdots,\ n-1)$，任何 $a_{kk}^{(k)}$ 均应非零，这样消去过程必可进行下去。由此我们便得到如下定理：

定理1.1：若 n 阶方阵 A 的逐次左上角主子式 $\det A_k\neq 0(k=1,\ 2,\ \cdots,\ n-1)$，则必可将其分解为下三角矩阵 L 与上三角矩阵 U 的乘积：

$$A=L\cdot U \tag{8.1.15}$$

系数矩阵的三角形分解定理在解线性代数方程组的直接法中起着重要的作用。除8.1.6节中的镜像映射法外，以后我们要介绍的直接解法大多数都以此为基础，仅是实现三角形分解的具体办法不同而已。读者根据这一点便很容易掌握这些方法之间的联系。

最后还应指出，矩阵 $A^{(n)}$ 对角线元素之乘积等于矩阵 $A^{(1)}$ 的行列式，即 $\det A^{(1)}=\prod_{k=1}^{n}a_{kk}^{(k)}$。所以，在消去过程中，系数矩阵 $A^{(1)}$ 的行列式是很容易求得的。但要注意在计算机上计算形如 $\prod_{k=1}^{n}a_{kk}^{(k)}$ 的连乘时很容易溢出(上溢或下溢)，必须适当地引入比例因子。

8.1.3 主元素消去法

（一）行（或列）交换的必要性

前述的消去法中，未知量是按其出现于方程式中的自然顺序消去的；用来消去其它方程式中未知量的方程亦是按顺序选取的，所以又叫做顺序消去法。实际计算已发现顺序消去的方式有很大的缺点。首先，消去过程中可能出现“主元素” $a_{kk}^{(k)}$（今后，我们约定把用作除数的元素 $a_{kk}^{(k)}$ 称为“主元素”）为零的情况，尽管 $\det \boldsymbol{A} \neq 0$，此时消去过程亦将无法进行下去。其次，如果“主元素” $a_{kk}^{(k)}$ 很小，由于舍入误差及有效位消失等因素，其本身常常有较大相对误差，用其作除数会导致其它元素数量级的严重增长和舍入误差扩散，使得最终解答极不准确。从下面的简单例子中，我们可以部分地看到上述情况。

考察如下二阶方程组：

$$\begin{cases} 0.0001\,x_1 + 1.00\,x_2 = 1.00 \\ 1.00\,x_1 \quad + 1.00\,x_2 = 2.00 \end{cases}$$

其真解舍入到小数点后五位之值为：$x_1 = 1.00010, x_2 = 0.99990$。假定我们采用三位十进制数的浮点计算，并用顺序消去法。第一步以 0.0001 为主元素，从第二个方程消去 x_1 后可得：

$$-10000\,x_2 = -10000,\ \text{从而有}\ x_2 = 1.00$$

回代之立即可得 $x_1 = 0.00$。显然，这一解答是无准确度可言的。得出这一错误结果的原因在于第一步的主元素太小，使得经消去后所得的三角形方程组很不准确所致。如果我们选取第二个方程式中 x_1 的系数 1.00 为主元素来消去第一个方程中的 x_1，则得出如下方程式：

$$1.00\,x_2 = 1.00$$

这样求得的解答为：$x_1 = 1.00,\ x_2 = 1.00$。显然，这是真解的三位正确舍入值。

从上述例子看出，消去过程中适当选取主元素是十分必要的。误差分析的理论结果和计算实践均表明，顺序消去法在系数矩阵 $\boldsymbol{A}$ 为对称正定时，可以保证计算过程对于舍入误差的数值稳定性，对一般的矩阵则必须引入适当选取主元素的技巧，方能得到满意的结果。

（二）主元素消去法

最常用和最有效的选取主元素的方法是列主元素消去法和全主元素消去法。

列主元素消去法中，未知数仍然是顺序地消去的，但是把各方程中要消去的那个未知数的系数中按模最大者作为主元素。即是每消去一个未知数（例如 x_k）时，在要消去的各方程中找出这个未知数系数中按模最大者（例如是第 i_k 个方程的系数 $a_{i_k k}^{(k)}$，$i_k \geqslant k$），并将第 i_k 个方程与第 k 个方程交换位置（此时 $a_{i_k k}^{(k)}$ 处于 $a_{kk}^{(k)}$ 的位置），再按顺序消去法的同样公式进行一步消去。交换两个方程的位置相当于在系数矩阵及自由项上都左乘以如下排列矩阵：

$$\boldsymbol{P}_k = \begin{bmatrix} 1 & & \vdots & & \vdots & \\ & \ddots\,1 & \vdots & & \vdots & \\ \cdots & \cdots & 0 & \cdots & 1 & \cdots \\ & & \vdots & 1 \ddots & \vdots & \\ \cdots & \cdots & 1 & \cdots & 0 & \cdots \\ & & \vdots & & \vdots & \end{bmatrix} \begin{matrix} \\ \\ k \\ \\ i_k \\ \\ \end{matrix} \tag{8.1.16}$$

所以，类似于(8.1.12)式，列主元素消去法所得出的三角形方程组应为：

$$A^{(n)}x=b^{(n)}$$

其中

$$A^{(n)}=M_{n-1}\cdot P_{n-1}\cdot M_{n-2}\cdot P_{n-2}\cdot\cdots\cdot M_2\cdot P_2\cdot M_1\cdot P_1\cdot A=L_1A \tag{8.1.17}$$

$$b^{(n)}=L_1b$$

解此三角形方程组即得出要求的解答。应当指出，如果 $\det A\neq 0$，从理论上说，列主元素消去法总是可以进行下去的。因为若经过选取主元素后仍有 $a_{kk}^{(k)}$ 为零，则矩阵第 k 列中从对角线元素起以下均应为零。故此列已消去完毕，可继续对第 $k+1$ 列进行消去，最后得到的三角形矩阵 $A^{(n)}$，其对角线上有零元素存在，故其行列式为零，亦即 $\det A$ 应为零，这与假设是矛盾的。所以，此时所有的 $a_{kk}^{(k)}$ 均不应为零。还应指出，列主元素法中矩阵 M_i 的元素 m_{ik} 之模必定小于等于1，因而消去过程中对舍入误差增长的控制较为有利。

全主元素消去法中，未知数不再按顺序消去。每步消去之前(例如第 k 步)，先在第 k 至第 n 个方程中的第 k 至第 n 个未知数的系数里，找出按模最大者作为主元素(即在系数矩阵的 k 到 n 行与 k 到 n 列中找出按模最大元素，例如是 $a_{i_kj_k}^{(k)}$)，然后将第 k 与第 i_k 个方程交换位置，将各方程中未知数 x_k 与 x_{j_k} 的两项交换位置(相当于系数矩阵及自由项的 k 行与 i_k 行，k 列与 j_k 列互换，此时，$a_{i_kj_k}^{(k)}$ 处于原来 $a_{kk}^{(k)}$ 的位置上)，再按顺序消去的公式进行一步消去。这样，逐次消去的未知数将由序列$\{j_k\}$决定。若令 Q_k 为如下排列矩阵：

$$Q_k=\begin{bmatrix} 1 & & & \vdots^{k} & & \vdots^{j_k} & & \\ & \ddots & & \vdots & & \vdots & & \\ & & 1 & \vdots & & \vdots & & \\ \cdots & \cdots & \cdots & 0 & \cdots & 1 & \cdots & \\ & & & \vdots & 1 & \vdots & & \\ & & & \vdots & \ddots & \vdots & & \\ & & & \vdots & 1 & \vdots & & \\ \cdots & \cdots & \cdots & 1 & \cdots & 0 & \cdots & \\ & & & \vdots & & \vdots & 1 & \\ & & & \vdots & & \vdots & & \ddots \\ & & & \vdots & & \vdots & & & 1 \end{bmatrix} \tag{8.1.18}$$

则全主元素消去法所得的三角形方程组可记为：

$$A^{(n)}y=b^{(n)}$$

其中

$$\begin{cases} A^{(n)}=M_{n-1}\cdot P_{n-1}\cdot M_{n-2}\cdot P_{n-2}\cdot\cdots\cdot M_2\cdot P_2\cdot M_1\cdot P_1\cdot A\cdot Q_1\cdot Q_2\cdot\cdots\cdot Q_{n-1}=L_2AQ \\ b^{(n)}=L_2\cdot b \\ y=Q^Tx \quad 或 \quad x=Qy \end{cases} \tag{8.1.19}$$

解此三角形方程组即得出 y。再顺次将 y 的第 $n-1$ 与 j_{n-1} 个分量交换，$n-2$ 个与 j_{n-2} 个分量交换，如此等等，直至最后，将 y 的第1个与第 j_1 个分量交换即得最终解答 x。与列主元素法类似，只要 $\det A\neq 0$，全主元素法在理论上总是可以进行下去的。并且，矩阵 M_i 的元素 m_{ik} 的模小于等于1，矩阵 $A^{(n)}$ 每行元素中以对角线元素按模最大，因而，消去及回代过程中对舍入误差增长的控制均较有利。还应指出，不论在列主元素法或全主元素法中，均有 $\det A=(-1)^s\cdot a_{11}^{(1)}\cdot a_{22}^{(2)}\cdots a_{nn}^{(n)}$，其中 s 为消去过程中行(或行和列)交换的总次数。因此消去过程中很容易附带地求出矩阵的行列式值。当然，也同样需要考虑比例因子问题以避免溢出。

列主元素消去法与全主元素消去法各有优缺点。一般说来，列主元素法已有足够的精

度，同时，由于全主元素法花费在寻找最大模元素及交换行列上的时间较多，也不能在消去过程中保持矩阵的某些有用的特点(例如带型或特殊的块型等)，程序也较复杂，所以经常使用列主元消去法。但全主元素消去法往往有更高的精确度，多花费一些机器时间有时还是值得的。因而，具体计算时应视问题的要求来选用合适的方法。

列主元素消去法和全主元素消去法解线性方程组的算法语言程序，见本章最后所附的程序1和程序2。

8.1.4 直接分解法

从系数矩阵 $\boldsymbol{A}$ 分解为下三角矩阵 $\boldsymbol{L}$ 与上三角矩阵 $\boldsymbol{U}$ 的乘积的定理1.1出发，我们可以直接得到计算 $\boldsymbol{L}$ 与 $\boldsymbol{U}$ 的元素的公式。这个公式不需要任何中间步骤，直接从矩阵 $\boldsymbol{A}$ 的元素算出 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 的元素，所以称之为直接分解法。求得 $\boldsymbol{L}$ 与 $\boldsymbol{U}$ 之后，便可将求解过程变为先解下三角形方程组 $\boldsymbol{Ly}=\boldsymbol{b}$，得出中间变量 $\boldsymbol{y}$，再用回代过程求解 $\boldsymbol{Ux}=\boldsymbol{y}$，得出最后结果 $\boldsymbol{x}$。把 $\boldsymbol{y}=\boldsymbol{Ux}$ 代入 $\boldsymbol{Ly}=\boldsymbol{b}$ 中便有：$\boldsymbol{LUx}=\boldsymbol{Ax}=\boldsymbol{b}$，故求得的 $\boldsymbol{x}$ 就是原方程组的解答。下面我们先以4阶矩阵为例说明其计算过程与存储单元的分配等等，然后再导出一般情况的计算公式。

考察如下4阶矩阵的分解式：

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \cdot \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

上面4阶矩阵等式等价于 4×4 个标量等式，但需要确定的 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 的元素共有20个，所以我们可令 $\boldsymbol{U}$ 的对角线元素为1，这样 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 的元素将唯一确定。

根据矩阵的乘法规则，$\boldsymbol{L}$ 的第一列元素显然等于 $\boldsymbol{A}$ 的相应元素(因 $u_{11}=1$)；矩阵 $\boldsymbol{U}$ 的第一行元素应为 $a_{1i}/l_{11}(i=2,\ 3,\ 4)$。为计算 $\boldsymbol{L}$ 的第二列元素，我们用 $\boldsymbol{L}$ 的第二至第四行分别乘 $\boldsymbol{U}$ 的第二列，即可得出：$l_{i2}=a_{i2}-l_{i1}\cdot u_{12}(i=2,\ 3,\ 4)$。由于 $\boldsymbol{U}$ 的第一行及 $\boldsymbol{L}$ 的第一列已算出，从此式便可直接求得矩阵 $\boldsymbol{L}$ 的第二列元素。再以 $\boldsymbol{L}$ 的第二行分别乘 $\boldsymbol{U}$ 的第三列，第四列，立即可得：$u_{2j}=(a_{2j}-l_{21}u_{1j})/l_{22}(j=3,\ 4)$，这样又可求出 $\boldsymbol{U}$ 之第二行。按同样方式，依次可以得出计算 $\boldsymbol{L}$ 第三列之公式为：$l_{i3}=a_{i3}-l_{i1}u_{13}-l_{i2}u_{23}\ (i=3,\ 4)$。计算 $\boldsymbol{U}$ 之第三行的公式则为：$u_{34}=(a_{34}-l_{31}u_{14}-l_{32}u_{24})/l_{33}$。最后，$l_{44}=a_{44}-l_{41}u_{14}-l_{42}u_{24}-l_{43}u_{34}$。这样就完成了矩阵 A 的三角形分解。

在计算机上求解时，通常需要把矩阵 $\boldsymbol{A}$ 及自由项 $\boldsymbol{b}$ 放在内存中。为了节省存储单元，求得的 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 的元素可以直接放在矩阵 $\boldsymbol{A}$ 相应元素的位置上，因为这时矩阵 $\boldsymbol{A}$ 的这些元素已经无用。这样，矩阵 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 就不再占用额外的存储单元。以上面4阶矩阵为例说明之。当求得 $\boldsymbol{L}$ 的第二列及 $\boldsymbol{U}$ 的第二行元素后，存储单元分配将如下所示：

$$\left[\begin{array}{cc:cc} l_{11} & u_{12} & u_{13} & u_{14} \\ l_{21} & l_{22} & u_{23} & u_{24} \\ \hdashline l_{31} & l_{32} & a_{33} & a_{34} \\ l_{41} & l_{42} & a_{43} & a_{44} \end{array}\right] \begin{bmatrix} b_1 \\ b_2 \\ \hdashline b_3 \\ b_4 \end{bmatrix}$$

可以看出上述计算过程中元素 l_{ii} 不能为零，否则计算无法进行下去。此外，若 l_{ii} 很小也将带来较大的舍入误差。所以，通常也采用与列主元消去法类似的办法进行处理。就是每求得一列 $\boldsymbol{L}$ 的元素后，找出该列中按模最大的元素 $l_{i_k k}$，然后将矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{L}$ 的第 i_k 行与第 k 行互换(此时 $l_{i_k k}$ 处于 l_{kk} 之位置)，再继续进行其后的计算。显然，这样做并不影响分解式中已经求得的 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 的其他元素，并且若同时将自由项 $\boldsymbol{b}$ 的第 i_k 个分量与第 k 个分量交换，亦不影响最终求得的解答。

综上所述，对于 n 阶矩阵 $\boldsymbol{A}$，直接分解法中计算 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 的元素的顺序为：第一步，先算 $\boldsymbol{L}$ 的第一列，然后 $\boldsymbol{U}$ 的第一行。第二步算 $\boldsymbol{L}$ 的第二列，然后 $\boldsymbol{U}$ 的第二行，如此等等，直至第 n 步算完 $\boldsymbol{L}$ 的第 n 列。注意：算完第 k 步时，$\boldsymbol{L}$ 的第 k 行即已求得，故可立即完成求解 $\boldsymbol{Ly}=\boldsymbol{b}$ 的第 k 步。所以，求解 $\boldsymbol{Ly}=\boldsymbol{b}$ 的过程也可以组合到上述各步中去。这样，我们便可把直接分解法的第 k 步计算过程归纳如下⊖：

(1) 按下列公式计算 $\boldsymbol{L}$ 的第 k 列：

$$l_{ik}=a_{ik}-\sum_{p=1}^{k-1} l_{ip}u_{pk} \quad (i=k,\ k+1,\ \cdots,\ n) \tag{8.1.20}$$

并将算出的 l_{ik} 存放于 a_{ik} 的位置上。

(2) 找出 l_{ik} 中按模最大元素，例如

$$|l_{i_k k}|=\max_{i\geq k}|l_{ik}|$$

然后交换存放矩阵 $\boldsymbol{A}$ 及自由项 $\boldsymbol{b}$ 的第 i_k 行与第 k 行的存储单元的内容(假定交换后原 a_{ij} 位置上的元素仍以符号 a_{ij} 来表示)。

(3) 按如下公式计算 $\boldsymbol{U}$ 的第 k 行元素：

$$u_{kj}=\left(a_{kj}-\sum_{p=1}^{k-1} l_{kp}\cdot u_{pj}\right)\Big/ l_{kk} \qquad (j=k+1, k+2,\ \cdots,\ n) \tag{8.1.21}$$

并将 u_{kj} 存放于 a_{kj} 的位置上。

(4) 按如下公式计算 $\tilde{b}_k$：

$$\tilde{b}_k=\left(b_k-\sum_{p=1}^{k-1} l_{kp}\tilde{b}_p\right)\Big/ l_{kk} \tag{8.1.22}$$

并将 $\tilde{b}_k$ 存于 b_k 位置上。

对 $k=1,\ 2,\ \cdots,\ n$，重复 n 次(1)～(4)的计算后，即求得方程式 $\boldsymbol{Ly}=\boldsymbol{b}$ 的解答 $\boldsymbol{y}$，并已存于 $\boldsymbol{b}$ 的位置上。再用公式(8.1.7)即可求得最终解答 $\boldsymbol{x}$。

如果不计寻找按模最大元素的工作量，第 k 步计算 l_{ik} 与 u_{kj} 分别需完成$(n-k+1)\cdot(k-1)$次乘法及加减法。故 n 步共需完成约 $2\sum(n-k)k\sim\frac{1}{3}n^3+O(n^2)$ 次乘法与加减法。所以直接分解法与消去法的计算工作量基本上是相同的。

从矩阵分解角度看，直接分解法与消去法本质上没有多大区别，但实际计算时它们各有长处。一般来说，如果仅用单字长进行运算，主元素消去法具有运算量较少、精度高的优点，故是最常用的。但是，为了提高精度往往采取单字长数双倍内积的办法(即作 $\sum_i a_ib_i$ 型计算时，采用双倍位加法，最终结果再舍入成单字长数)，这时直接分解法是最适宜的，理论分析与计算实践均表明，一般来说它能够获得较高的精度(参见[11], pp. 65; [18])。

直接分解法的算法语言程序见本章最后所附的程序 3。由于双倍内积运算不能用算法

⊖ 注意：本章里我们约定，当求和号 $\sum$ 的上限小于下限时，该和为零。

语言表示，该程序内我们仍使用单字长运算。同时，为了反复使用方便起见，在程序中先将 $\boldsymbol{A}$ 分解，然后再分别解 $\boldsymbol{Ly}=\boldsymbol{b}$ 及 $\boldsymbol{Ux}=\boldsymbol{y}$，即把前述计算步骤(4)移至分解后进行。

8.1.5 对称正定矩阵的平方根法和 LDL^T 分解法

对称正定矩阵在实践中经常遇到，由于其本身的特点，使用前述的解法是不利的，应该采用适合其特点的解法。这里介绍的平方根法和 $\boldsymbol{LDL}^T$ 分解法，其运算量和存储量较普通消去法均节省一半左右，而且不需要选主元素，求得解答的精度也很高，是目前在计算机上解这类问题最有效的方法之一。

假设我们要解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中 $\boldsymbol{A}$ 为实对称正定矩阵。由于 $\boldsymbol{A}^T=\boldsymbol{A}$，若 $\boldsymbol{A}$ 的某个实三角形分解式为 $\boldsymbol{A}=\tilde{\boldsymbol{L}}\cdot\tilde{\boldsymbol{U}}$，则必有 $\tilde{\boldsymbol{U}}^T\cdot\tilde{\boldsymbol{L}}^T=\tilde{\boldsymbol{L}}\cdot\tilde{\boldsymbol{U}}$，或者 $\tilde{\boldsymbol{L}}^{-1}\cdot\tilde{\boldsymbol{U}}^T=\tilde{\boldsymbol{U}}\cdot\tilde{\boldsymbol{L}}^{-T}$。最后一个等式的左边为下三角阵，右边为上三角阵，所以两者均应为对角阵 $\boldsymbol{D}$。这样便有：$\tilde{\boldsymbol{L}}=\boldsymbol{D}\cdot\tilde{\boldsymbol{L}}^T$，或者 $\boldsymbol{A}=\tilde{\boldsymbol{L}}\boldsymbol{D}\tilde{\boldsymbol{L}}^T$。再从 $\boldsymbol{A}$ 的正定性得知其任一左上角主子式 $\boldsymbol{A}_k$ 均为正。对于等式 $\boldsymbol{A}_k=\tilde{\boldsymbol{L}}_k\cdot\boldsymbol{D}_k\cdot\tilde{\boldsymbol{L}}_k$ 两端求行列式便可得知 $\boldsymbol{D}$ 的对角线元素均为正数。这样一来，令 $\boldsymbol{L}=\tilde{\boldsymbol{L}}\cdot\boldsymbol{D}^{1/2}$，便可证明在矩阵 $\boldsymbol{A}$ 的三角形分解式(8.1.15)中可有 $\boldsymbol{U}=\boldsymbol{L}^T$，即分解式可以写为：

$$\boldsymbol{A}=\boldsymbol{L}\cdot\boldsymbol{L}^T \tag{8.1.23}$$

从这一等式出发，按直接分解法类似的步骤，我们很容易得出计算矩阵 $\boldsymbol{L}$ 的元素的公式。这里与直接分解法不同的是矩阵 $\boldsymbol{U}=\boldsymbol{L}^T$，故 $\boldsymbol{U}$ 的对角线元素不再等于1。此外在求得 $\boldsymbol{L}$ 的第 k 列元素之后，$\boldsymbol{L}^T$ 的第 k 行元素即已得出，所以其计算量将为直接分解法的一半左右，即 $\frac{1}{6}n^3+O(n^2)$ 次乘法与加减法。同时从(8.1.23)式我们有：$a_{ii}=\sum\limits_{j=1}^{i}l_{ij}^2$，从矩阵 $\boldsymbol{A}$ 的正定性又得知 $\det\boldsymbol{A}_i>0$，于是可以证明：$\max\limits_{i,j}\{l_{ij}^2\}\leqslant\max\limits_{i}\{a_{ii}\}$，并且 $l_{ii}>0$。所以分解过程中各元素 l_{ij} 的数量级不会增长，对角线元亦恒为正数，这样选主元素就不必要了。计算实践也表明不选主元素已有足够精度。因而，我们只要顺序地计算下三角形矩阵 $\boldsymbol{L}$ 的第一列至第 n 列的元素即可。按照直接分解法的计算公式(8.1.20)和(8.1.21)并利用 $u_{kj}=l_{jk}$，读者很容易推出矩阵 $\boldsymbol{L}$ 第 k 列元素的计算公式为：

$$\begin{cases} l_{kk}=\left(a_{kk}-\sum\limits_{p=1}^{k-1}l_{kp}^2\right)^{\frac{1}{2}} \\ l_{jk}=\left(a_{jk}-\sum\limits_{p=1}^{k-1}l_{jp}l_{kp}\right)\Big/l_{kk}\quad(j=k+1,\ k+2,\ \cdots,\ n) \end{cases} \tag{8.1.24}$$
$$(k=1,\ 2,\ \cdots,\ n)$$

这个计算公式中只用到矩阵 $\boldsymbol{A}$ 下三角部分的元素，算得的 l_{ij} 同样也可以存放在 a_{ij} 的位置上。所以，只需在计算机的内存中留出 $n(n+1)/2$ 个单元即可进行计算。这样，平方根法所需存储量也约为消去法的一半左右。

如果矩阵 $\boldsymbol{A}$ 对称但不是正定矩阵，上述分解式仍然成立，不过此时 l_{ii} 可能成为虚数。此外，为了控制舍入误差的增长，也必须引入主元素，这样将破坏对称性。所以，在这种情况下仍以采用某种主元素消去法为宜。

平方根法的算法语言程序见本章最后所附的程序4，该程序仍然采用先进行分解，然后解三角形方程组的办法，以利于反复多次使用。

用平方根法解对称正定方程组时，要完成 n 次开方运算。绝大多数计算机上开方运算

是用子程序实现的，这样将增加不少运算量，其结果精确度有时也稍差。我们可以设法避免开方运算来解决这个问题。

从(8.1.24)知道，若将下三角矩阵 $\boldsymbol{L}$ 的第 k 列元素提出公因子 l_{kk} 便得到:

$$\boldsymbol{L}=\tilde{\boldsymbol{L}}\cdot\boldsymbol{D}=\begin{bmatrix}1 & & & & \\ \tilde{l}_{21} & 1 & & \boldsymbol{0} & \\ \tilde{l}_{31} & \tilde{l}_{32} & 1 & & \\ \vdots & \vdots & & \ddots 1 & \\ \tilde{l}_{n1} & \tilde{l}_{n2} & \cdots & \tilde{l}_{n,n-1} & 1\end{bmatrix}\cdot\begin{bmatrix}l_{11} & & & \\ & l_{22} & & \boldsymbol{0} \\ & & \ddots & \\ \boldsymbol{0} & & & l_{nn}\end{bmatrix}$$

这样一来，分解式(8.1.23)便可写成:

$$\boldsymbol{A}=\tilde{\boldsymbol{L}}\boldsymbol{D}^2\tilde{\boldsymbol{L}}^T$$

如果用 $\tilde{\boldsymbol{D}}$ 表示 $\boldsymbol{D}^2$，我们便得到对称正定矩阵 $\boldsymbol{A}$ 的下列分解式:

$$\boldsymbol{A}=\tilde{\boldsymbol{L}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{L}}^T \tag{8.1.25}$$

其中，$\tilde{\boldsymbol{L}}=[\tilde{l}_{ij}]$为单位下三角阵，并且 $\tilde{l}_{ij}=l_{ij}/l_{jj}$; $\tilde{\boldsymbol{D}}$ 为有正对角线元素 $\tilde{d}_i=l_{ii}^2$ 的对角型矩阵。

从(8.1.24)很容易推得计算分解式(8.1.25)中矩阵 $\tilde{\boldsymbol{L}}$ 与 $\tilde{\boldsymbol{D}}$ 的元素 $\tilde{l}_{jk}$ 和 $\tilde{d}_k$ 的公式为:

$$\begin{cases}\tilde{d}_k=a_{kk}-\sum\limits_{p=1}^{k-1}\tilde{l}_{kp}^2\cdot\tilde{d}_p \\ \tilde{l}_{jk}=\left(a_{jk}-\sum\limits_{p=1}^{k-1}\tilde{l}_{jp}\cdot\tilde{l}_{kp}\cdot\tilde{d}_p\right)\Big/\tilde{d}_k \quad (j=k+1,\ k+2,\ \cdots,\ n)\end{cases}$$
$$(k=1,\ 2,\ \cdots,\ n)$$

为了节省乘法运算，可令 $\tilde{a}_{jk}=\tilde{l}_{jk}\cdot\tilde{d}_k$，并将计算次序改为按行计算 $\tilde{\boldsymbol{L}}$ 的元素㊀，从上式即可得如下计算公式:

$$\begin{cases}\tilde{a}_{jk}=a_{jk}-\sum\limits_{p=1}^{k-1}\tilde{a}_{jp}\cdot\tilde{l}_{kp},\ \tilde{l}_{jk}=\tilde{a}_{jk}/\tilde{d}_k \quad (k=1,\ 2,\ \cdots,\ j-1) \\ \tilde{d}_j=a_{jj}-\sum\limits_{p=1}^{j-1}\tilde{a}_{jp}\cdot\tilde{l}_{jp}\end{cases} \tag{8.1.26}$$
$$(j=1,\ 2,\ \cdots,\ n)$$

容易看出，这一公式所需完成的乘法量约为 $\frac{1}{6}n^3$，但没有开方运算。

求得矩阵 $\tilde{\boldsymbol{L}}$ 和 $\tilde{\boldsymbol{D}}$ 后，解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 即可分三步完成:

(1) 解下三角形方程组 $\tilde{\boldsymbol{L}}\boldsymbol{y}=\boldsymbol{b}$，得出向量 $\boldsymbol{y}$。

(2) 计算向量 $\boldsymbol{z}$ 的各分量:

$$z_i=y_i/\tilde{d}_i \quad (i=1,\ 2,\ \cdots,\ n)$$

(3) 解上三角形方程组 $\tilde{\boldsymbol{L}}^T\boldsymbol{x}=\boldsymbol{z}$，即得最终解答 $\boldsymbol{x}$。

为便于使用起见，我们在本章最后所附的程序 5 中给出按这个方法编制的算法语言程序。

8.1.6 镜像映射法

前面讨论的消去法、直接分解法和平方根法本质上都是某种实现系数矩阵三角形分解的办法。这里我们要介绍另外一种将系数矩阵分解为正交矩阵和上三角矩阵乘积的方法，

㊀ 这样作主要是为了节省存放 $\tilde{a}_{jk}$ 的存储单元，因为计算 $\tilde{\boldsymbol{L}}$ 的第 j 行元素时只用到 $\tilde{a}_{jp}(p=1,\ 2,\ \cdots,\ j-1)$，前面的 $\tilde{a}_{pq}$ 就不需要保存(详见本章最后的程序 5)。

即镜像映射法或称 Householder 方法。这个方法除了用于解线性方程组外，在最小二乘问题(见 8.3 节)和代数特征值问题(见本书第十章)中还有广泛的应用。

在 8.1.2 节中曾经看到，高斯消去法实际上是一种用一系列初等变换矩阵 $\boldsymbol{M}_i$ 左乘方程组的系数矩阵和自由项，从而将其化为等价三角形方程组的方法。所以又叫作用初等变换的三角化过程。除了用初等变换矩阵外，还可以采用正交矩阵来进行三角化，并获得很高的精确度。镜像映射矩阵就是其中最有效的一种。为说明方便起见，我们先介绍关于镜像映射矩阵的基本概念，然后讨论如何用它来进行三角化和完成正交-三角分解。

(一)镜像映射矩阵

镜像映射矩阵(又称初等埃尔米特矩阵或 Householder 矩阵)是因其几何意义得名的。为说明这一点，我们先从普通三维空间中的几何关系谈起。

如果我们将三维空间一向量 $\boldsymbol{z}$ 对平面 Q 作镜像映射(即找出向量 $\boldsymbol{z}$ 对于镜平面 Q 的"像")，便得到向量 $\boldsymbol{z}^*$。容易看出(见图 8.1)有如下关系：

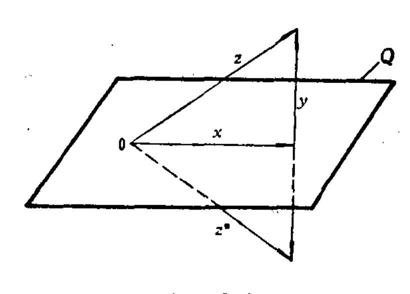


图 8.1

$$\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{y}, \quad \boldsymbol{z}^*=\boldsymbol{x}-\boldsymbol{y}$$

其中向量 $\boldsymbol{x}$ 属于平面 Q，向量 $\boldsymbol{y}$ 与 Q 正交。

以 $\boldsymbol{w}$ 表示平面 Q 的单位法向量(即 $\boldsymbol{w}^T\cdot\boldsymbol{w}=1$)，则：

$$\begin{cases} \boldsymbol{w}^T\cdot\boldsymbol{x}=0 \\ \boldsymbol{y}=\alpha\boldsymbol{w} \quad \text{或者} \quad \boldsymbol{w}=\dfrac{1}{\rho}(\boldsymbol{z}-\boldsymbol{z}^*) \end{cases} \tag{8.1.27}$$

其中，α 和 ρ 为常数，其值分别等于向量 $\boldsymbol{y}$ 和 $\boldsymbol{z}-\boldsymbol{z}^*$ 的长度。

由 $\boldsymbol{z}$ 至 $\boldsymbol{z}^*$ 的变换可以用矩阵形式表示，我们令：

$$\boldsymbol{H}=\boldsymbol{I}-2\boldsymbol{w}\cdot\boldsymbol{w}^T \tag{8.1.28}$$

则有：

$$\begin{aligned}\boldsymbol{H}\cdot\boldsymbol{z}&=(\boldsymbol{I}-2\boldsymbol{w}\cdot\boldsymbol{w}^T)\cdot\boldsymbol{z}=\boldsymbol{z}-2\boldsymbol{w}\cdot(\boldsymbol{w}^T\cdot\boldsymbol{z})=\boldsymbol{x}+\alpha\boldsymbol{w}-2\boldsymbol{w}\cdot[\boldsymbol{w}^T\cdot(\boldsymbol{x}+\alpha\boldsymbol{w})]\\&=\boldsymbol{x}+\alpha\boldsymbol{w}-2\alpha\boldsymbol{w}=\boldsymbol{x}-\alpha\boldsymbol{w}=\boldsymbol{z}^*\end{aligned}$$

所以，矩阵 $\boldsymbol{H}$ 代表将空间任意向量对平面 Q 作镜像映射的线性变换，故矩阵 $\boldsymbol{H}$ 称为镜像映射矩阵。对于一般的 n 维空间，(8.1.28)所定义的矩阵 $\boldsymbol{H}$ 也有类似的几何解释。因而我们就以(8.1.28)作为镜像映射矩阵的定义。容易看出，矩阵 $\boldsymbol{H}$ 不仅是对称矩阵，而且也是正交矩阵。因为：

$$\begin{aligned}\boldsymbol{H}\cdot\boldsymbol{H}^T&=(\boldsymbol{I}-2\boldsymbol{w}\cdot\boldsymbol{w}^T)\cdot(\boldsymbol{I}-2\boldsymbol{w}\cdot\boldsymbol{w}^T)=\boldsymbol{I}-4\boldsymbol{w}\cdot\boldsymbol{w}^T+4\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{w})\boldsymbol{w}^T\\&=\boldsymbol{I}-4\boldsymbol{w}\cdot\boldsymbol{w}^T+4\boldsymbol{w}\cdot\boldsymbol{w}^T=\boldsymbol{I}\end{aligned}$$

镜像映射矩阵 $\boldsymbol{H}$ 由单位向量 $\boldsymbol{w}$ 唯一确定。恰当地选择向量 $\boldsymbol{w}$，可使相应矩阵 $\boldsymbol{H}$ 乘空

间任意向量 s 后得出与任意单位向量 l 平行的向量 $\alpha\cdot l$。为此只需将 s 看作 z, αl 看作 z^*, 所以可令:

$$w=\frac{1}{\rho}(s-\alpha l) \tag{8.1.29}$$

其中

$$\alpha=-\operatorname{sign}(s^T\cdot l)\cdot\sqrt{s^T\cdot s}$$

$$\rho=\sqrt{(s-\alpha l)^T\cdot(s-\alpha l)}=\sqrt{2\alpha^2-2\alpha(s^T\cdot l)}=\sqrt{2(\alpha^2+|\alpha|\cdot|s^T\cdot l|)}$$

此时便有:

$$H\cdot s=s-2w\cdot w^T\cdot s=s-\frac{2}{\rho}w(s^T-\alpha l^T)\cdot s=s-\frac{2}{\rho}w(\alpha^2+|\alpha|\cdot|s^T\cdot l|)$$
$$=s-\rho w=\alpha l$$

应该指出,因为正交变换下向量的长度不变,所以 $|\alpha|$ 应等于向量 s 的长度。至于 α 的符号的选取,主要是使得计算 ρ 时不会因为两数相减而造成有效位数消失。

矩阵 H 的上述特性今后要经常用到。

(二)三角化过程和正交-三角分解

利用(8.1.28)和(8.1.29),很容易构造出 $n-1$ 个镜像映射矩阵 $H_i(i=1, 2, \cdots, n-1)$, 来达到三角化的目的。其大体过程如下:首先,我们取矩阵 $A^{(1)}=A$ 的第一列:$(a_{11}, a_{21}, a_{31}, \cdots, a_{n1})^T$ 为 s, 取单位向量 $e_1=(1, 0, 0, \cdots, 0)^T$ 为 l, 按(8.1.28)和(8.1.29)来形成变换矩阵 H_1, 则容易得知:

$$A^{(2)}=H_1\cdot A^{(1)}=\begin{bmatrix} \alpha & a_{12}^{(2)} & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & \cdots & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & \cdots & a_{nn}^{(2)} \end{bmatrix}$$

如果再取 $A^{(2)}$ 的第二列(令第一个元素为0)为 s, 即 $(0, a_{22}^{(2)}, a_{32}^{(2)}, \cdots, a_{n2}^{(2)})^T=s$, 取 $e_2=(0, 1, 0, \cdots, 0)^T$ 为 l, 形成变换矩阵 H_2, 用 H_2 左乘 $A^{(2)}$, 同样可将 $A^{(2)}$ 第二列主对角线以下元素化为零,并保持矩阵 $A^{(2)}$ 的第一列及第一行不变。此时变换后矩阵的前两列对角线以下元素已变为零。类似地,取 $A^{(k)}$ 的第 k 列(令前 $k-1$ 个元素为零)为 s, 即 $(0, \cdots, 0, a_{kk}^{(k)}, \cdots, a_{nk}^{(k)})^T=s$, 取 $e_k=(0, \cdots, 0, 1, 0, \cdots, 0)^T$ 为 l, 形成矩阵 H_k, 则 $H_kA^{(k)}$ 的前 $k-1$ 行和 $k-1$ 列将保持不变,而其第 k 列对角线以下元素变为零。依此类推,作 $n-1$ 次乘法后,矩阵 A 将被化为上三角形矩阵 R:

$$R=A^{(n)}=H_{n-1}\cdot H_{n-2}\cdot\cdots\cdot H_2\cdot H_1\cdot A=\begin{bmatrix} \times & \times & \cdots & \times \\ & \times & \cdots & \times \\ & & \times \ddots & \times \\ \mathbf{0} & & \ddots & \vdots \\ & & & \times \end{bmatrix} \tag{8.1.30}$$

如果对自由项 b 也同时左乘以逐次的矩阵 H_i, 显然,我们将得到与(8.1.2)等价的三角形方程组:

$$A^{(n)}x=b^{(n)} \tag{8.1.31}$$

其中，$b^{(n)}=H_{n-1}\cdot H_{n-2}\cdot\cdots\cdot H_2\cdot H_1\cdot b$。用回代过程(8.1.7)解此上三角形方程组即得最终解答。以上就是用镜像映射矩阵求解方程组的计算过程。

实际进行 $H\cdot A$ 型的运算时，不需将矩阵 H 明显地求出来然后作矩阵乘法，只需用 H 去乘 A 的各列得出 $H\cdot A$ 的相应列。在作 H 乘 A 的列时，用如下公式计算：

$$H\cdot A_i=A_i-2(w^T\cdot A_i)\cdot w \tag{8.1.32}$$

每算一列只需作一个向量内积，一个数量乘向量和一个向量加法，这对于减少工作量是有意义的。

按照(8.1.28)、(8.1.29)，很容易推得用镜像映射矩阵进行三角化的计算公式为：

$$\begin{cases} H_k=I-2w_kw_k^T \quad A^{(k+1)}=H_k\cdot A^{(k)} \quad b^{(k+1)}=H_k\cdot b^{(k)} \\ w_k=u_k/\rho_k \\ \rho_k=[2\alpha_k(\alpha_k+|a_{kk}^{(k)}|)]^{1/2} \\ \alpha_k=\left[\sum_{l=k}^{n}(a_{lk}^{(k)})^2\right]^{1/2} \\ u_k=(0,\ 0,\ \cdots,\ 0,\ a_{kk}^{(k)}+\operatorname{sign}(a_{kk}^{(k)})\cdot\alpha_k,\ a_{k+1,k}^{(k)},\cdots,\ a_{nk}^{(k)})^T \end{cases} \tag{8.1.33}$$

$$(k=1,\ 2,\ \cdots,\ n-1)$$

实际计算时，为了节省运算量常将上述公式按(8.1.32)的形式稍加变形。由于

$$A^{(k+1)}=H_k\cdot A^{(k)}=A^{(k)}-2w_k\cdot w_k^T\cdot A^{(k)}=A^{(k)}-\frac{u_k}{\rho_k^2}(2u_k^TA^{(k)})$$

若令 $q_k^T=2u_k^T\cdot A^{(k)}/\rho_k^2$，则有：

$$A^{(k+1)}=A^{(k)}-u_k\cdot q_k^T$$

完全类似地有

$$b^{(k+1)}=b^{(k)}-(2u_k^T\cdot b^{(k)}/\rho_k)\cdot u_k$$

这样一来，把需要计算的各量按计算的先后次序排列，便得到如下计算 $A^{(n)}$ 及 $b^{(n)}$ 的公式：

$$\begin{cases} \alpha_k=\left[\sum_{l=k}^{n}(a_{lk}^{(k)})^2\right]^{1/2} \\ u_k=(0,\ \cdots,\ 0,\ a_{kk}^{(k)}+\operatorname{sign}(a_{kk}^{(k)})\alpha_k,\ a_{k+1,k}^{(k)},\cdots,\ a_{nk}^{(k)})^T \\ \sigma_k=2\alpha_k\cdot(\alpha_k+|a_{kk}^{(k)}|) \\ q_k^T=2u_k^TA^{(k)}/\sigma_k \\ \mu_k=2u_k^T\cdot b^{(k)}/\sigma_k \\ A^{(k+1)}=A^{(k)}-u_k\cdot q_k^T \\ b^{(k+1)}=b^{(k)}-\mu_ku_k \end{cases} \tag{8.1.34}$$

$$(k=1,\ 2,\ \cdots,\ n-1)$$

从上述公式可知，第 k 步计算约需完成 $2\times[(n-k+1)^2+(n-k+1)]$ 次乘法和加减法运算。所以三角化过程总共约需完成 $\frac{2}{3}n^3+O(n^2)$ 次乘法和加减法运算及 n 次开方运算。与高斯消去法相比较，镜像映射法的运算量多一倍左右。正是由于这个原因，尽管从精确度方面来说消去法有时可能略差一些，但实践中却使用得广泛得多。但是，矛盾方程组的最小二乘解问题则是例外。由于该类问题的特点(参见 8.3 节)，采用镜像映射法更为优越一些。

镜像映射法解线性方程组的算法语言程序见本章最后所附的程序 6。

最后，我们讨论一下正交-三角分解问题。利用关系式 $H_i=H_i^T=H_i^{-1}$，从(8.1.30)式我

们得知 $A=(H_1\cdot H_2\cdot\cdots\cdot H_{n-2}\cdot H_{n-1})\cdot R=QR$。其中 $Q=H_1\cdot H_2\cdot\cdots\cdot H_{n-1}$ 为正交矩阵。这就是说，任意矩阵 A 总可以分解为正交矩阵 Q 与上三角形矩阵 R 的乘积。如果 A 是非奇矩阵，我们还可以说明这一分解式在相差一个对角线元素之模为 1 的对角矩阵 D 之下是唯一的。实际上，若有两种分解：

$$A=Q_1R_1=Q_2R_2$$

由于 A 非奇，所以 R_1 也是非奇异的，这样我们有：

$$Q_2^TQ_1=R_2R_1^{-1}$$

上式左端说明矩阵$(R_2R_1^{-1})$为正交矩阵，但它同时应是上三角形矩阵。因而，它只能是对角线型矩阵，其对角线元之模必为1(实数情况必为±1)。这就是说：

$$Q_2^TQ_1=R_2R_1^{-1}=D=\begin{bmatrix} e^{i\theta_1} & & & \mathbf{0} \\ & e^{i\theta_2} & & \\ & & \ddots & \\ \mathbf{0} & & & e^{i\theta_n} \end{bmatrix}$$

或者

$$Q_1=Q_2D,\quad R_1=D^{-1}\cdot R_2$$

综上所述，我们有如下正交-三角分解定理：

定理 1.2：

任意实(复)矩阵 A，总可分解为正交(酉-)矩阵 Q 与上三角矩阵 R 的乘积：

$$A=Q\cdot R$$

如果 A 是非奇异的，则 Q 之各列与 R 之各行在相差一个模为 1 的常数下唯一确定。这一定理在本书第十章中讨论代数特征值问题的数值解时将要用到。

8.1.7 求逆矩阵问题

有些计算问题中常常需要求出给定矩阵 A 的逆矩阵。例如，结构分析问题中有时需要求出刚度矩阵的逆矩阵，回归分析问题中需要求出相关矩阵的逆矩阵，线性规划问题和某些非线性方程组数值求解问题等，有时也需要求出某些矩阵的逆矩阵。因而，有必要讨论一下计算机上常用的求逆矩阵方法。

求逆矩阵的问题本质上与解线性方程组问题是相同的。如果要求矩阵 A 的逆矩阵 X，从关系式 $AX=I$ 出发，容易验证，我们只需要解如下 n 个线性方程组：$AX_i=e_i(i=1, 2, \cdots, n)$，其中 e_i 为第 i 个分量为 1；其他分量为零的单位向量；X_i 即为 A^{-1} 的第 i 列。但是，这个办法有很大缺点。首先，除去存放 A 的 n^2 个存储单元外，还需要 n^2 个存放 A^{-1} 的存储单元。其次，它没有利用自由项的特殊形状以节省所需运算量。但这些缺点是很容易克服的，例如只要把前面所述的消去法稍加变形就可得到比较有效的算法。我们先来讨论这个方法。

在顺序消去法的第 k 步，我们只消去了系数矩阵中第 k 列对角线以下的元素，这样得出的等价方程组的系数矩阵是上三角形矩阵 U。如果我们在计算的第 k 步也同时将第 k 列的对角线以上的元素消去，并使系数矩阵第 k 列变为单位向量 e_k，这样，消去完毕后，系数矩阵就将变成单位矩阵 I。完全类似于8.1.2节中的讨论，我们容易得知这样的消去过程也相当于对系数矩阵逐次左乘以一个初等变换矩阵。用矩阵符号表示便有：

$$\overline{M}_n \cdot \overline{M}_{n-1} \cdot \cdots \cdot \overline{M}_3 \cdot \overline{M}_2 \cdot \overline{M}_1 \cdot A = I \tag{8.1.35}$$

其中

$$\overline{M}_k = \begin{bmatrix} 1 & & m_{1k}^{(k)} & & \\ & \ddots & \vdots & & 0 \\ & & 1 \quad m_{kk}^{(k)} & & \\ & & \vdots & 1 & \\ & 0 & \vdots & & \ddots \\ & & m_{nk}^{(k)} & & 1 \end{bmatrix} \qquad m_{ik}^{(k)} = \begin{cases} -\dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, & i \neq k \\ \dfrac{1}{a_{kk}^{(k)}}, & i = k \end{cases} \tag{8.1.36}$$

将(8.1.35)式两端右乘以 A^{-1}，根据逆矩阵的定义，我们就有：

$$A^{-1} = \overline{M}_n \cdot \overline{M}_{n-1} \cdot \cdots \cdot \overline{M}_2 \cdot \overline{M}_1 \tag{8.1.37}$$

于是我们得出了逆矩阵 A^{-1} 的因式分解式，通常称之为逆矩阵 A^{-1} 的乘积形式。这一消去过程就称为高斯-若当消去法。

因式矩阵 $\overline{M}_k$ 与单位矩阵不同之处仅在第 k 列，因而只需 n 个存储单元即可保存它。再注意到经第 k 步消去后，原矩阵第 k 列已变为单位向量 e_k，其后的计算中这些存储单元不再使用。所以，我们可将 $\overline{M}_k$ 的第 k 列元素存放在原矩阵的第 k 列位置上。这样，保存逐次的因式矩阵 $\overline{M}_k$ 将不占用额外存储单元。此外，从(8.1.37)可以看出，将 $\overline{M}_k$ 按消去过程的顺序连乘起来，就得出 A^{-1}。如果逐次地将 $\overline{M}_k$ 的第 k 列存于被消去后矩阵的第 k 列位置上，下一步消去时，再以 $\overline{M}_{k+1}$ 左乘此整个矩阵的每一列，容易验证，这时已经存放了 $\overline{M}_1$，$\overline{M}_2$，…，$\overline{M}_k$ 的各列位置上元素之变化，与作(8.1.37)中的连乘积完全相同。所以，也就自然地完成了求 $\overline{M}_{k+1} \cdot \overline{M}_k \cdot \cdots \cdot \overline{M}_2 \cdot \overline{M}_1$ 的连乘运算。消去过程结束时，矩阵 A 的位置上即为要求的结果 A^{-1}。于是，整个求逆过程中，除去存放矩阵 A 的 n^2 个单元外，不再需要其它存储单元。这一存储安排格式通常叫做矩阵的原地求逆格式。

综上所述，用高斯-若当消去法，计算逆矩阵的第 k 步，可以归纳如下(这里，我们用 A_{ij} 表示当时处于系数矩阵第 i 行和第 j 列位置上的元素)：

(1) 计算 $a = 1/A_{kk}$，并送到 A_{kk} 所在位置。

(2) 对于 $i \neq k$，计算 $-aA_{ik}$，并送到 A_{ik} 所在位置。

(3) 对于 $i \neq k$，$j \neq k$，计算 $A_{ij} + A_{ik} \cdot A_{kj}$，并送到 A_{ij} 所在位置。

(4) 对于 $j \neq k$，计算 aA_{kj}，并送到 A_{kj} 所在位置。

很容易看出，第 k 步需完成 $(n+1)(n-1)$ 次乘法，$(n-1)^2$ 次加法。所以，n 步消去共需完成 $n(n^2-1)$ 次乘法及 $n(n-1)^2$ 次加法，即高斯-若当消去法求逆矩阵所需的乘法和加法量均为 n^3 量级，比之用消去法解一个方程组来说，大约是其三倍。由于这个原因，先求出逆矩阵 A^{-1}，再去计算 $A^{-1} \cdot b$ 来作为方程组的解答是不合算的，即使在多个自由项时，如果保留矩阵 A 的三角分解式来求解，一般也比保留 A^{-1} 合算。因此，除指定必需求出逆矩阵的情况外，一般都不采用求逆矩阵的办法来解线性方程组。

如果矩阵 A 是对称正定矩阵，A^{-1} 也应是对称正定的，并且，计算 A^{-1} 的过程中也有对称的性质可以利用。这样，我们可以在内存中仅存放原始矩阵的上三角形部分，并将逆矩阵的上三角形部分也存在同一位置上，以达到节省存储单元的目的。按照这一存储格式编制的对称正定矩阵原地求逆程序见本章最后所附的程序 7。

和解方程组的消去法一样，为了保证求逆结果的精确度，对于非对称正定的矩阵，引入选主元素技巧是十分必要的。通常也采用列主元素或全主元素法，其过程与解方程组时完

全类似。例如，对于全主元素法，我们将有：

$$\overline{M}_n\cdot P_n\cdots\overline{M}_2P_2\overline{M}_1P_1\cdot A\cdot Q_1\cdot Q_2\cdots Q_n=I$$

或者

$$A^{-1}=Q_1Q_2\cdots Q_n\cdot\overline{M}_n\cdot P_n\cdot\overline{M}_{n-1}\cdot P_{n-1}\cdots\overline{M}_2\cdot P_2\cdot\overline{M}_1\cdot P_1 \tag{8.1.38}$$

矩阵 P_k 与 Q_k 均为形如(8.1.16)和(8.1.18)的初等排列矩阵，只需一个行号(或列号) $i_k(j_k)$ 即可表示出来。因而，我们用 $2n$ 个存储单元记录相应的 i_k 与 $j_k(k=1, 2, \cdots, n)$ 后，仍可采用前述的原地求逆存储格式。这样，全主元素法求逆矩阵过程的第 k 步可以归纳如下(A_{ij} 的意义同前)：

(1) 找出 $\max\limits_{k<i,j<n}|A_{ij}|$ 的位置(i_k, j_k)，并记录 i_k, j_k。

(2) 交换整个矩阵的第 k 行与 i_k 行，k 列与 j_k 列。

(3) 计算 $\alpha=1/A_{kk}$，并送至 A_{kk} 所在位置。

(4) 对于 $i\neq k$，计算 $-\alpha\cdot A_{ik}$，并送至 A_{ik} 所在位置。

(5) 对于 $i\neq k$ 及 $j\neq k$，计算 $A_{ij}+A_{ik}\cdot A_{kj}$，并送至 A_{ij} 所在位置。

(6) 对于 $j\neq k$，计算 $\alpha\cdot A_{kj}$，并送至 A_{kj} 所在位置。

对于 $k=1, 2, \cdots, n$，执行上述计算后，矩阵 A 的位置上就存放着逆矩阵 A^{-1} 的元素。但由于计算过程中我们进行了行列互换，因而，A^{-1} 的元素已不按其应有的行列次序排列。为了得出正确次序的结果，必须再进行相应的行列互换工作。在前面原地求逆的计算过程与存储安排中，已经存放在前 k 列的矩阵 $\overline{M}_kP_k\cdots\overline{M}_2P_2\overline{M}_1P_1$ 的不同于单位阵的 k 个列已进行了行交换 $i_{k+1}\to k+1$，而连乘时 $\overline{M}_{k+1}$ 之元素本应存于 i_{k+1} 列，但我们将它存于 $k+1$ 列位置上。所以，可以知道，此时矩阵 A 的位置上所存放的矩阵与 $\overline{M}_n\cdot P_n\cdots\overline{M}_1P_1$ 仅差一个列的排列 $i_k\to k(k=n, n-1, \cdots, 1)$，再从(8.1.38)式得知最后还应进行如下行列交换：

(1) 第 i_k 列与 k 列互换$(k=n, n-1, \cdots, 2, 1)$。

(2) 第 j_k 行与 k 行互换$(k=n, n-1, \cdots, 2, 1)$。

按上述方案编制的全主元素法求逆矩阵程序见本章最后所附的程序8。

我们还要指出，用镜像映射法求逆矩阵时，也很容易按原地求逆的存储格式来安排。因为由(8.1.30)有：

$$A^{-1}=R^{-1}H_{n-1}\cdot H_{n-2}\cdots H_1 \tag{8.1.39}$$

从(8.1.34)知，逐次变换矩阵 H_i 仅需 $n-i+1$ 个存储单元即可保留，故可存于矩阵第 i 列对角线以下的位置上(注意！α_k 为 $A^{(k+1)}$ 的第 k 个主对角线元，故另需 n 个工作单元来存放 $a_{kk}^{(k)}$)。上三角矩阵 R 的求逆很容易在原地进行。最后逐次右乘 H_i 的过程也可在矩阵右下角 $n-i+1$ 阶子式的位置上完成，故最终在原始矩阵 A 的位置上即可得到逆矩阵 A^{-1}。根据这一安排，读者不难自行编出相应的算法语言程序。

8.1.8 特殊形状矩阵和高阶矩阵问题的直接解法

前面讨论的解方程组 $Ax=b$ 的方法是适用于阶数不高的所谓"满矩阵"的，即用于矩阵的绝大多数元素皆为非零的情况。然而实践中大量遇到的矩阵并不是低阶的"满矩阵"，一般说来，它们的阶数较高(例如，几百阶或上千阶)，并且总具有某些特殊形状。例如在结构分析问题，大地测量问题，电力传输网分析问题以及各种常微或偏微分方程数值解问题等方面，经常需要求解一个高阶的线性代数方程组，其系数矩阵一般具有下列几种形状：

(i) 带状矩阵或变带宽带状矩阵。它们满足条件：

$$a_{ij}=0 \quad 若 \quad i>j+m_1 \quad 或 \quad i<j-m_2$$

其形状如图 8.2 所示：

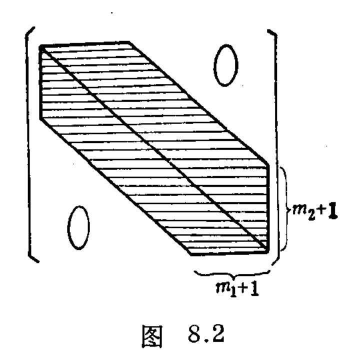


图 8.2

其中，m_1 为主对角线下面部分的"带宽"；m_2 为上面部分的"带宽"；m_1+m_2+1 为该带型矩阵的"总带宽"。如果带的内部各行(列)的非零元素的"宽度"不等，矩阵将呈如图 8.3 所示形状：

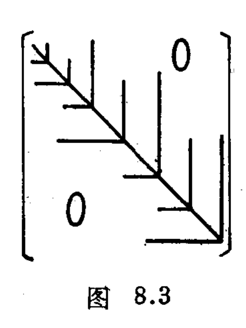


图 8.3

则叫作变带宽带状矩阵。

(ii) 加边带状矩阵，即上述带状矩阵加上一个宽度为 s 的边，其形状如图 8.4 所示：

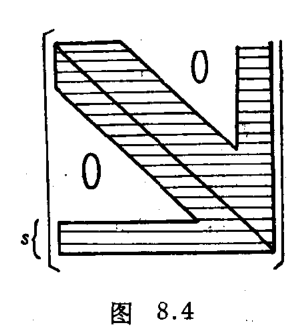


图 8.4

(iii) 加边对角块矩阵。即分块对角型矩阵加上一个宽度为 s 的边，形状如图 8.5 所示：

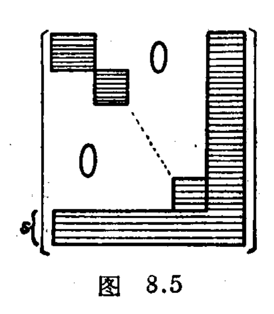


图 8.5

(iv) 块三对角型矩阵，形状如图 8.6 所示：

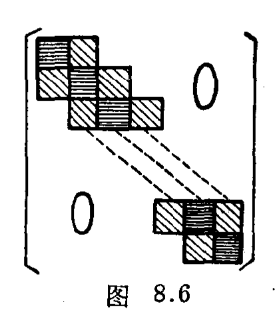


图 8.6

(v) 某种绝大多数元素为零，但非零元素不规则分布的矩阵。

上述这些矩阵的共同特点是它们都有大量的零元素。实际问题中常常是非零元素仅占 5～10% 左右，有的问题中非零元素仅占 1% 或更少一些。这样的矩阵，通常称为"稀疏矩阵"或"稀矩阵"。利用零元素很多和按一定规则分布等特点，可以大大地节省计算工作量和有效地提高在计算机上能解问题的阶数。实践中常用的解法都是针对这些特点而提出的特殊方法。因而，我们在处理问题时应按照具体问题具体分析的原则，采用适当方法。下面我们讨论处理这类问题的几种常用方法。

(一)矩阵分解法或消去法

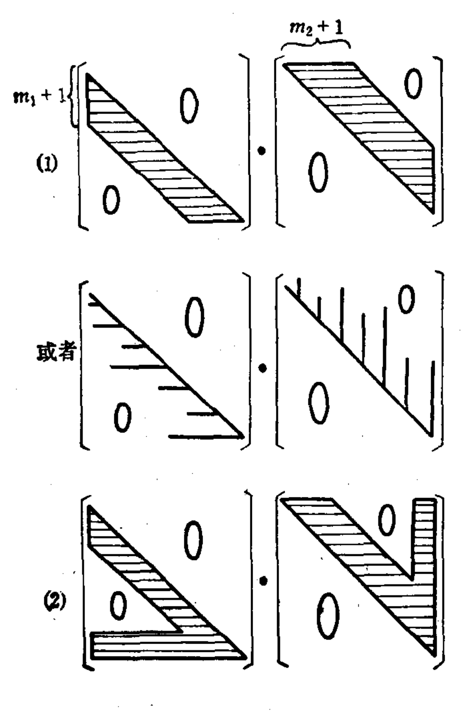
首先我们考虑直接分解法，从矩阵的三角形分解过程可以看出，若矩阵 A 的某行(例如第 i 行元素)满足条件：

$$a_{ij}=0, \quad 1\leqslant j\leqslant k_i<i$$

即是说第 i 行的前 k_i 个元素都为零。当我们按计算 l_{ij} 的公式(8.1.20)进行计算时，可以验证相应地也有：

$$l_{ij}=0, \quad 1\leqslant j\leqslant k_i<i$$

这样，对上述(i)～(iv)中的矩阵来说，矩阵 L 与矩阵 A 的下三角部分应该有相同的"形状"(我们这里假定不引入选主元素技巧，并对"形状"两字作轮廓地理解)。同样推理可知，矩阵 U 与矩阵 A 的上三角部分亦有相同"形状"。因此，上面所列举的前四种矩阵其相应的分解式应如图 8.7 所示：



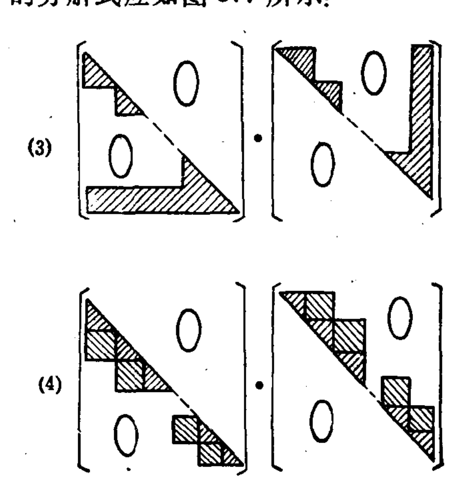


图 8.7

对于第(v)类矩阵，前述结论是不成立的，因其非零元素分布并无一定的规则外形，在分解与消去过程中，原来矩阵中某些零元素的位置将出现新的非零元素，矩阵 $\boldsymbol{L}$ 与 $\boldsymbol{U}$ 的非零元素分布将与 $\boldsymbol{A}$ 的相应部分有所不同。

实际计算中，我们可以利用(i)～(iv)型矩阵的分解式的特点，在进行分解或消去时，仅计算其中非零部分的元素，这样就可以大大节省计算量与存储量。以带状矩阵为例，若按上述原则进行处理，其顺序消去过程所需的加法和乘法量分别为$m^2\cdot n$(这里假设$m_1=m_2=m$). 所需的存储量大致为$(2m+1)\cdot n$。显然，当 $m\ll n$, 较满阵时的 $n^3/3$ 次加法与乘法以及 n^2 个存储单元是有本质上的节省的。因而，目前在计算机上大都采用这个原则来解带状方程组。

由于要利用矩阵的特殊形状来进行计算，就必须对不同类型的矩阵编制相应的计算程序，这些程序的编制原则是很类似的。下面我们仅以带状矩阵为例来说明程序的编制原则，并按两种情形分别处理：

(1) 对称正定情况

处理带状矩阵时，如果矩阵是对称正定的，同样不需要选主元素，结果精度也能得到保证。因而直接采用平方根法的分解过程，按照矩阵的特点编制出仅计算其分解式中非零元素的程序，即可有效地解决问题。下面我们来说明这一程序的编制方法。

假设矩阵 $\boldsymbol{A}$ 对称正定且为带状，总带宽为$(2m+1)$，所以其分解式 $\boldsymbol{A}=\boldsymbol{L}\boldsymbol{L}^T$ 中的 $\boldsymbol{L}^T$也有与 $\boldsymbol{A}$ 的上三角部分相同的带状形式。在存储安排上，我们只将矩阵 $\boldsymbol{A}$ 上三角部分的非零元素按行存放在内存中作为原始数据。例如存于场 $A[1:n,\ 1:m+1]$内，其具体排列如下：

$$\begin{bmatrix} a_{11},\ a_{12},\ \cdots,\ a_{1,m+1} \\ a_{22},\ a_{23},\ \cdots,\ a_{2,m+2} \\ \cdots\cdots \\ a_{n-m,n-m},\cdots,\ a_{n-m,n} \\ \cdots\cdots \\ a_{n-1,n-1},\ a_{n-1,n},\ \times,\cdots,\ \times \\ a_{n,n},\ \times,\ \cdots,\ \times \end{bmatrix} \tag{8.1.40}$$

这样矩阵 $\boldsymbol{A}$ 在"带"以外的零元素就不占用存储单元。同时，由于 $\boldsymbol{L}^T$ 的每行非零元素也为 $m+1$ 个，并且在算出这些元素后，矩阵 $\boldsymbol{A}$ 的相应元素已不再需要，所以，作为计算结果，我们又可将 $\boldsymbol{L}^T$ 的元素逐行地存入场 A 内的相应位置上。求得 $\boldsymbol{L}^T$ 后，解方程组 $\boldsymbol{L}\boldsymbol{y}=\boldsymbol{b}$ 及 $\boldsymbol{L}^T\boldsymbol{x}=\boldsymbol{y}$ 是很容易实现的。这样，求得解答的过程中，我们无需涉及带外的任何元素。按照这个安排编制的算法语言程序见本章最后所附的程序 9。读者参照前述的平方根法程序(程序 4)和上述安排，将会比较快地理解和掌握它。

对于对称正定的变带宽矩阵，存储格式要稍为改变一下。如果第 i 行第一个非零元素的列号为 j_i, 则我们仅存储该行的第 j_i 至第 i 个元素，并按行顺序地将矩阵的这些元素排列起来，构成一个一维场 D。每行(例如 i 行)对角线元素在场 D 中的位置，将另外用一个一维场 S 记录下来，即 $S[i]$代表第 i 行对角线元素在场 D 中的位置。于是，矩阵第 i 行 j 列处的元素在场 D 中的位置应是 $S[i]-i+j$。所以我们很容易找到它。按照这个安排和考虑到分解式中下三角矩阵 $\boldsymbol{L}$ 的第 i 行非零元仅出现在第 j_i 列至第 i 列位置上，便很容易编制出相应的解方程组程序，其细节我们不再赘述。应该指出，这一处理办法虽然程序略为复杂

些，但有时能比带状矩阵的处理办法获得好得多的效果。因而，是值得推荐的。这个方法的计算公式及算法语言程序见本章最后所附的程序10及其说明。

还应指出，有些带状矩阵虽不是对称正定的，但其具有较强的对角优势条件，这时也无需引入选主元素的技巧。例如，偏微分方程初值或边值问题的数值求解中经常遇到的某些三对角线方程组即属此种情况，其系数矩阵 $\boldsymbol{A}$ 具有下列形状：

$$\boldsymbol{A}=\begin{bmatrix} a_1 & b_1 & & & & \\ c_2 & a_2 & b_2 & & & \\ & c_3 & a_3 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & & c_n & a_n \end{bmatrix} \tag{8.1.41}$$

其中元素 a_i, b_i, c_i 满足所谓对角优势条件：

$$\begin{cases} \text{(i)}\ |a_1|>|b_1|>0 \\ \text{(ii)}\ |a_i|\geqslant|c_i|+|b_i|,\ \text{且}\ b_i\cdot c_i\neq 0 \quad (i=2,\ 3,\ \cdots,\ n-1) \\ \text{(iii)}\ |a_n|>|c_n|>0 \end{cases} \tag{8.1.42}$$

假定上述矩阵 $\boldsymbol{A}$ 的三角形分解式为：

$$\boldsymbol{A}=\begin{bmatrix} \alpha_1 & & & \mathbf{0} \\ c_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ \mathbf{0} & & c_n & \alpha_n \end{bmatrix}\cdot\begin{bmatrix} 1 & \beta_1 & & \\ & 1 & \beta_2 & \\ & & \ddots & \ddots \\ & & & \ddots\ \ \beta_{n-1} \\ & & & 1 \end{bmatrix}$$

并把方程组的自由项 $\boldsymbol{b}$ 记为：

$$\boldsymbol{b}^T=(f_1,\ f_2,\ \cdots,\ f_n)$$

按照前述直接分解法不选主元素的计算公式 (8.1.20)、(8.1.21)、(8.1.22) 以及 (8.1.7)，就能推得如下计算公式：

$$\begin{cases} \beta_1=a_1^{-1}\cdot b_1,\quad v_1=a_1^{-1}\cdot f_1,\quad \alpha_1=a_1 \\ \alpha_i=a_i-c_i\cdot\beta_{i-1}, \\ \beta_i=\alpha_i^{-1}\cdot b_i, \qquad\qquad (i=2,\ 3,\ \cdots,\ n) \\ v_i=\alpha_i^{-1}\cdot(f_i-c_i v_{i-1}), \end{cases}$$
$$\begin{cases} x_n=v_n, \\ x_i=v_i-\beta_i\cdot x_{i+1}, \quad (i=n-1,\ n-2,\ \cdots,\ 1) \end{cases} \tag{8.1.43}$$

这就是通常称之为“追赶法”的计算公式。在对角优势条件 (8.1.42) 满足时，显然 $|\beta_1|<1$；同时，若 $|\beta_{i-1}|<1$ 则有：

$$|\beta_i|\leqslant|a_i-c_i\beta_{i-1}|^{-1}\cdot|b_i|\leqslant||a\,|-|c_i|\cdot|\beta_{i-1}||^{-1}\cdot|b_i|<(|a_i|-|c_i|)^{-1}\cdot|b_i|\leqslant 1$$

所以，所有 β_i 之模均应小于1。这样一来，就有：

$$|\alpha_i|>|a_i|-|c_i|\geqslant|b_i|>0$$

因而，计算过程必定可以进行下去，并且不会出现中间结果数量级的巨大增长和相应的舍入误差的严重发展。更仔细的分析表明，计算过程每步所引入的舍入误差将在其后的各步中

逐步减少，即是说，这一计算公式对于舍入误差是稳定的(参见[2]第338页)。

用追赶法解三对角方程组的算法语言程序见本章最后所附的程序11。

还应指出，如果矩阵 $\boldsymbol{A}$ 是所谓分块三对角线矩阵，即(8.1.41)中的 $\boldsymbol{a}_i$、$\boldsymbol{b}_i$、$\boldsymbol{c}_i$ 均为子矩阵，且 $\boldsymbol{a}_i$ 为方阵时，若将向量 $\boldsymbol{x}$ 及自由项 $\boldsymbol{b}$ 亦作相应分块，并用 $\boldsymbol{x}_i$ 和 $\boldsymbol{f}_i$ 来记这些子向量，那么，公式(8.1.43)仍是正确的。同时，容易证明只要矩阵 $\boldsymbol{A}$ 的逐次分块主子矩阵 $\boldsymbol{A}^{(k)}$ 之行列式非零，(8.1.43)对于矩阵 $\boldsymbol{A}$ 总可以进行下去。自然，在计算 $\boldsymbol{a}_i^{-1}\cdot\boldsymbol{b}_i$ 等各项时，我们宁可采取解线性方程组的办法而避免求出逆矩阵 $\boldsymbol{a}_i^{-1}$。

(2) 非对称情况

如果遇到的特殊形状矩阵是非对称的或者是对称但非正定的，一般说来，为了控制舍入误差的增长，采取某种选主元素的技巧是必要的。这时，使用全主元素法可能导致完全破坏矩阵的特殊形状，所以，一般都采用列主元素法。自然，用列主元素法进行消去时，矩阵的特殊形状亦将受到一定破坏，但只要我们仔细分析计算过程，仍可找出如何利用其特殊形状来节省运算量和存储量的规律。下面我们以非对称带状矩阵为例来说明这一点。

考虑形如前面所列举的(i)的非对称带状矩阵。假设用列主元素消去法(即进行行交换)来解相应的方程组。容易看出，从逐次的主列(例如第 k 列)的非零元素(指在对角线元及紧靠其下面的 m_1 个元素)中选出主元素，并将其所在之行(称为主行)与第 k 行交换，然后用它消去其它行(例如第 i 行，$k<i\leqslant k+m_1$)中第 k 列处的非零元素时，被消去的行中主对角线右边非零元素的个数可能增加，但其增加的个数最多不超过 m_1 个(当 $k+m_1$ 行为主行时，将可能增加 m_1 个)。因此，若不计被消去的非零元素，整个消去过程中矩阵各行非零元素之和不超过 m_1+m_2+1 个(实际上，大多数情况下每消去一步所涉及的各行主对角线元左边有一个元素变为零，右边增加一个非零元素)。这样，我们就可以用 $n\times(m_1+m_2+1)$ 个存储单元来存放矩阵 $\boldsymbol{A}$ 和逐次消去过程中的中间矩阵，并安排为如下形式(以8阶矩阵为例，其中 $m_1=2,\ m_2=3$)：

$$\begin{bmatrix} \times & \times & a_{11} & a_{12} & a_{13} & a_{14} \\ \times & a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} \\ a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} \\ a_{64} & a_{65} & a_{66} & a_{67} & a_{68} & \times \\ a_{75} & a_{76} & a_{77} & a_{78} & \times & \times \\ a_{86} & a_{87} & a_{88} & \times & \times & \times \end{bmatrix}$$

消去开始时，首先将前两行改排为：

$$\begin{bmatrix} a_{11}, & a_{12}, & a_{13}, & a_{14}, & 0, & 0 \\ a_{21}, & a_{22}, & a_{23}, & a_{24}, & a_{25}, & 0 \end{bmatrix}$$

再在前三行的第一个元素中选主元，并将主行换至第一行位置，进行消去。然后再将第二、三行改排为：

$$\begin{bmatrix} a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} & a_{26}^{(1)} & 0 \\ a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} & a_{35}^{(1)} & a_{36}^{(1)} & 0 \end{bmatrix}$$

并在第二、三、四行上重复上述消去过程，如此进行，直至消完。这时，所得上三角方程组的

系数矩阵即占据原来存放矩阵 $\boldsymbol{A}$ 的位置，其对角线元在第一列上。回代过程(解 $\boldsymbol{Ux}=\tilde{\boldsymbol{b}}$)是显然的。按这一过程编制的算法语言程序见本章最后所附的程序 12。

(二)分块消去法

实践中经常出现计算机的内存容纳不下矩阵 $\boldsymbol{A}$ 全部非零元素的情况，需要使用外存(磁鼓或磁盘)来存储所需数据。这种情况在中、小型计算机或是在多道程序的大型计算机上(其内存仅有指定的一部分为某个题目使用)解高阶方程组时经常遇到，这时采取分块消去法是较为适宜的。

所谓分块消去法，就是用某种方法将矩阵分成若干子块，并将这些子块看作元素，按照普通消去法的格式进行相应的子块运算来求得方程组的解答。这样，我们就可以把全部矩阵元素按子块为单位存放在计算机的外存中，而仅将当前计算所要用到的子块调到内存中去参加运算。由于外存容量一般总是比内存大得多，所以这种安排能够提高可以求解问题的阶数。

显然应该怎样进行分块和组织内外存的数据交换才能取得较好的效果，是首先必须解决的问题。读者不难看出，内存中留作交换数据用的单元数 M 愈多，效果将愈好。除此之外，计算格式的适当选取也是很重要的因素。下面我们对于"满矩阵"介绍一种常用的而且比较有效的格式。

假设内存可用于存放子块的单元数为 M，令 $P=\left[\sqrt{\frac{M}{4}}\right]\left(\text{即}\sqrt{\frac{M}{4}}\text{的整数部分}\right)$。我们假定矩阵和自由项(列数$\leqslant P$)按如下形式进行分块:

$$
[\boldsymbol{A},\ \boldsymbol{b}]=\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r} & \boldsymbol{b}_1 \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r} & \boldsymbol{b}_2 \\ \cdots\cdots \\ \boldsymbol{A}_{r1} & \boldsymbol{A}_{r2} & \cdots & \boldsymbol{A}_{rr} & \boldsymbol{b}_r \end{bmatrix}
$$

其中除 $\boldsymbol{A}_{ir}$ 及 $\boldsymbol{A}_{rj}$ 以外，其他 $\boldsymbol{A}_{ij}$ 都是 $P\times P$ 方阵，而 $\boldsymbol{A}_{rr}$ 为 $S\times S$ 方阵($S\leqslant P$)，并且 $P(r-1)+S=n$。这样，内存中就可以同时容纳四个子块，分块消去过程可按下述步骤进行:

(1) 从外存调进子块 $\boldsymbol{A}_{ii}$，并原地进行三角形分解，得出 $\boldsymbol{A}_{ii}=\boldsymbol{L}\cdot\boldsymbol{U}$。

(2) 从外存调进子块 $\boldsymbol{A}_{i,i+k}$，用求解三角形方程组的办法算出 $\boldsymbol{A}_{ii}^{-1}\cdot\boldsymbol{A}_{i,i+k}$，保存在$\boldsymbol{A}_{i,i+k}$位置上，同时将其送至外存 $\boldsymbol{A}_{i,i+k}$ 位置上。

(3) 从外存调进子块 $\boldsymbol{A}_{ji}$，算出 $\boldsymbol{A}_{ji}\cdot\boldsymbol{A}_{ii}^{-1}\cdot\boldsymbol{A}_{i,i+k}$，并保存在内存中存放 $\boldsymbol{A}_{ji}$ 的位置上。

(4) 从外存调进子块 $\boldsymbol{A}_{j,i+k}$，算出 $\boldsymbol{A}_{j,i+k}-\boldsymbol{A}_{ji}\cdot\boldsymbol{A}_{ii}^{-1}\cdot\boldsymbol{A}_{i,i+k}$ 并送至外存$\boldsymbol{A}_{j,i+k}$的位置上。

(5) 对于 $j=1,\ 2,\ \cdots,\ r\ (j\neq i)$，重复步骤 (3) 与 (4)。

(6) 对于 $k=1,\ 2,\ \cdots,\ r-i+1$，重复 (2)～(5)(我们假设 $\boldsymbol{A}_{i,r+1}=\boldsymbol{b}_i$)。

(7) 对于 $i=1,\ 2,\ \cdots,\ n$，重复(1)～(6)。

最终解答处于自由项 $\boldsymbol{b}$ 的位置上。

前述格式亦可改为只需内存同时容纳三个子块的格式，此时 $\boldsymbol{L}\cdot\boldsymbol{U}$ 分解不予保留，(6)中相应改为重复(1)～(5)。这一格式节省了内存需要量，但却增加了计算量，在内存较小时是可以采用的。

从上述过程可以看出，使用外存的分块消去法比矩阵全部放在内存的通常消去法需要更多的计算时间，这是由于要完成许多额外的内外存数据交换工作。如果我们以 T 表示数

据交换所需要的总时间，T_R 表示分块消去法真正的计算时间，T_0 表示整个矩阵均放在内存时通常消去法所需时间，并将如下数值：

$$P=\frac{T_R+T}{T_0}$$

叫做时间增长系数，那么，显然，P 值愈大，利用外存的分块计算效果愈差。为了得出 P 值大小的一些估计，有人在某些计算机上对不同情况（即不同计算格式和不同的 M 和 n 值）作过实验，其结果表明（参见[5]），上述 P 值一般将小于 2。并且随着 M 与 n 的增大 P 值将单调下降。自然，随着外存方面的技术进步，这一数值还可以减小。不过，由此已可看出，目前在有足够容量外存（磁鼓、磁盘）的中、小型计算机上，完全可以用分块消去法有效地解高阶问题。

分块消去法的另一优点就是可以充分利用矩阵的特殊分块形状，进一步节省运算量和提高可以解决的问题的阶数。例如，对于本节开始所列举的两种特殊分块矩阵(iii)和(iv)，使用分块消去法将获得显著效果。下面我们分别讨论一下这两种特殊分块矩阵的消去过程。

对于分块三对角矩阵，处理的办法与普通带状矩阵完全类似。也就是说，我们仅将自由项列和系数矩阵中的非零子块（按行的顺序）存放于外存中。进行消去时，每次可从外存调若干个（例如 r 个）块行至内存（自然块行数 r 根据内存大小而定）。然后，将这些块行中对角线以下子块消去，这时，该 r 行矩阵块的变化如下面所示：

$$\begin{bmatrix} \boldsymbol{A}_{kk}^{(k)}, & \boldsymbol{A}_{k,k+1} & & & & \boldsymbol{b}_k^* \\ \boldsymbol{A}_{k+1,k}, & \boldsymbol{A}_{k+1,k+1}, & \boldsymbol{A}_{k+1,k+2} & & & \boldsymbol{b}_{k+1} \\ & \boldsymbol{A}_{k+2,k+1}, & \boldsymbol{A}_{k+2,k+2}, & \boldsymbol{A}_{k+2,k+3} & & \boldsymbol{b}_{k+2} \\ & \ddots & \ddots & \ddots & & \vdots \\ & & \boldsymbol{A}_{k+r,k+r-1} & \boldsymbol{A}_{k+r,k+r} & \boldsymbol{A}_{k+r,k+r+1} & \boldsymbol{b}_{k+r} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B}_{k,k+1} & & & & \boldsymbol{b}_k^* \\ 0 & \boldsymbol{I} & \boldsymbol{B}_{k+1,k+2} & & & \boldsymbol{b}_{k+1}^* \\ & \ddots & \ddots & \ddots & & \vdots \\ & 0 & \boldsymbol{I} & & \boldsymbol{B}_{k+r-1,k+r} & \boldsymbol{b}_{k+r-1}^* \\ & & 0 & \boldsymbol{A}_{k+r,k+r}^{(k+r)}, & \boldsymbol{A}_{k+r,k+r+1} & \boldsymbol{b}_{k+r}^* \end{bmatrix}$$

子块 $\boldsymbol{B}_{i,i+1}$ 及 $\boldsymbol{b}_i^*(i=k, k+1, \cdots, k+r-1)$ 在以后的消去中不再使用，我们可将其存放至外存中去。此时 $\boldsymbol{A}_{k+r,k+r}^{(k+r)}$ 已处于 $\boldsymbol{A}_{kk}^{(k)}$ 相同的地位。故可将第 $k+r$ 行移至原第 k 行位置上，再从外存调进 r 个块行，重复上述过程，直至消去过程完毕。回代时，只需从外存顺次调进 $\boldsymbol{B}_{i,i+1}$ 及 $\boldsymbol{b}_i^*$，解答即可求得。读者不难看出，整个计算过程中只有原来矩阵中的非零子块参加运算和占用存储单元，因此，运算量和存储量都获得较大节约。特别是在许多实际问题中，矩阵的子块是由程序根据某些初始数据计算出来的，这时，我们就不需要事先把它们算出来存放在外存中，而只需在用到它们时由程序临时产生，整个计算过程则仅需保存子块 $\boldsymbol{B}_{i,i+1}$。这样，可以更进一步节省存储单元。如果矩阵是分块五对角型或一般地 $2m+1$ 块对角型的，处理的办法也完全类似，我们不再赘述。

对于加边块对角型矩阵，利用分块消去过程在节省存储量方面将更为有利。为了讨论方便，我们把矩阵写为：

$$[A, b]=\begin{bmatrix} A_{11} & & 0 & A_{1r} & b_1 \\ & A_{22} & & A_{2r} & b_2 \\ & & \ddots & \vdots & \vdots \\ & 0 & & A_{r-1,r-1} & A_{r-1,r} & b_{r-1} \\ A_{r1} & A_{r2} & \cdots A_{r,r-1} & A_{rr} & b_r \end{bmatrix}$$

为了将第 r 块行中的非对角子块消去，只需顺次用第一块行消去 A_{r1}，第二块行消去 A_{r2}，如此等等。最终即可得出下列矩阵：

$$\begin{bmatrix} I & & & A_{11}^{-1}\cdot A_{1r} & A_{11}^{-1}\cdot b_1 \\ & I & & A_{22}^{-1}\cdot A_{2r} & A_{22}^{-1}\cdot b_2 \\ & & \ddots & \vdots & \vdots \\ & & I & A_{r-1,r-1}^{-1}\cdot A_{r-1,r} & A_{r-1,r-1}^{-1}\cdot b_{r-1} \\ 0 & 0 & \cdots 0 & \left(A_{rr}-\sum_{i=1}^{r-1} A_{ri}\cdot A_{ii}^{-1}\cdot A_{ir}\right) & b_r^* \end{bmatrix}$$

回代过程是显然的。上述消去过程较之三对角块消去过程不同之处主要在于需要保存的中间子块 $A_{ii}^{-1}\cdot A_{ir}$ 等很容易直接从原始矩阵的相应子块中求得。如果矩阵的子块是由程序产生的，我们就可以不保留子块 $A_{ii}^{-1}\cdot A_{ir}$，而在回代过程中用到它们时临时去产生。这样虽然浪费了一些计算时间，但却大大地节省了存储量。采用这一方案，真正需要自始至终存储的仅是上式中最右下角的子块$\left(A_{rr}-\sum_{i=1}^{r-1} A_{ri}\cdot A_{ii}^{-1}\cdot A_{ir}\right)$。在许多实际问题中这一子块也有特殊形状（如三对角块或五对角块等等），利用其特点还可以将所需存储量进一步减少。因而，在一般中小型计算机上用这一办法可以解决许多阶数很高的问题。

最后还应指出，计算 $A_{ii}^{-1}\cdot A_{ir}$ 时，最好使用解方程组 $A_{ii}X=A_{ir}$ 的办法。因为 A_{ii} 往往是特殊形状的矩阵（如带状稀疏的），解方程组比直接求逆无论在计算量或存储量上都要节省得多。解决具体问题时，如何通过行列的调换把矩阵划分成加边块对角型形式是应用上述方法的关键。一般来说，应使得 A_{rr} 的阶数尽可能低，或者使得$\left(A_{rr}-\sum_{i=1}^{r-1} A_{ri}\cdot A_{ii}^{-1}\cdot A_{ir}\right)$具有某种特殊形状，这对于存储与计算量的节省有重要的作用。最好是结合问题的物理背景来划分，这样做一般都可取得较好的效果。例如结构分析中的子结构方法，电网分析中的网络分割法等等，都是利用各自问题中的特点来实现上述的分块和消去过程的例子。由于这些方法简单有效，在结构分析和电网分析等等计算中是经常使用的。

（三）稀疏矩阵技术

对于某些非零元素分布不规则的稀疏矩阵（例如，晶体管电路分析问题中所遇到的矩阵），采用带状矩阵分解法或特殊分块形状的消去法是难以充分利用矩阵的稀疏性质的。这时，可以采用数据去零存储技术和逻辑尺（或下标表）定位的办法进行处理。这就是说我们仅将矩阵的非零元素依次存放起来，并用一个逻辑尺（或下标表）来标明这些元素的位置（逻辑尺共 n^2 个二进制位，是由 $[n^2/p]+1$ 个存储单元组成的，其中 n 为矩阵阶数，p 为计算机的字长。逻辑尺的 n^2 位中每个二进制位对应于一个矩阵的元素，若该元素非零，则该位为1，否则为零）。在求解过程中，通过分析和修改逻辑尺（或下标表）来判定哪些中间结果需要计算和存储，以节省与零元素相应的那些不必要的运算。这样就可达到节约计算工作量与存储量的目的。

下面我们用消去法为例,说明这一办法的基本思想。在消去过程中,原始矩阵零元素位置上一般将出现一些新的非零元素。消去的次序不同,新产生的非零元素的个数也很不相同。为了使求解过程尽可能有效,必须适当地重新排列方程与未知数的次序,使中间过程产生的非零元素尽可能少。最简单的排次序方法是把消去时不使其它行产生新的非零元素的行排在前面,然后将剩下的行按其中非零元素个数的多少顺次排列(一般说,非零元素少的排在前面将使其它行产生较少的新的非零元素),同样地对列的顺序也作类似处理。这种简单办法可以取得一定效果,但是比较粗略。还有许多更有效的办法,由于比较繁琐,在此不一一叙述,有兴趣的读者可以参见[24]中有关文章。

排完次序后,即可按通常顺序消去过程进行计算。为使零元素不参加运算,并记录新产生的非零元素,我们将反复使用逻辑尺来判定主行应该与哪些行进行运算,其非零元素乘以常数后应与哪行相应元素相加(若该元素不为零),或直接送至该元素位置上并修改逻辑尺(该元素为零,此时将产生一个新的非零元素)等等。这样逐行进行下去,最终即得出上三角形矩阵 $\boldsymbol{U}$ 的全部非零元素及相应逻辑尺。自由项也在上述过程中相应改变。再利用新的逻辑尺(对应于 $\boldsymbol{U}$ 的部分)进行去零回代,求解过程即告结束。如果要反复求解具有同一系数矩阵的方程组,我们可以只分析一遍逻辑尺,将求得的 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 连同相应的逻辑尺一道存放起来。以后求解时只需求解去零的下三角形和上三角形方程组。如果要反复求解同样稀性结构但非零元素之值每次不同的方程组,也可用一面分析逻辑尺,一面产生相应的计算程序的办法来解决。一旦程序产生后,反复求解过程就不需要再分析和修改逻辑尺了。

对于非常稀疏的矩阵,使用逻辑尺可能会浪费存储单元(因为逻辑尺总需要 $[n^2/p]+1$ 个单元)。这时可用下标表来处理,即用一系列单元来存放非零元素的下标(即行号或列号)。计算过程中,反复分析和修改这个表以达到去零的目的,其过程与用逻辑尺的办法大致类似。其细节读者可参阅[24]中有关文章。

上述这些就是去零存储技术的基本思想。由于其细节过于繁琐,我们不再详细讨论。但是我们要指出,这一方法有效性的关键在于巧妙地排次序和使用逻辑尺或下标表。关于这两个问题已有许多专门的研究,其中有的方法对某些问题已经取得显著效果。例如,在晶体管开关电路分析问题中,需要重复求解稀疏性结构相同的线代数方程组,大都按上述这类方法进行处理,并已编制了一系列的分析程序。这些程序解算一次方程组的时间,可以降至通常满矩阵方法的几十分之一,所需的运算量也从满矩阵的 $n^3/3$ 降至 $O(n)$ 的数量级。所以,仔细地排列方程与未知数的次序和适当地选取逻辑尺或下标的构造与加工方法,再加上充分利用反复求解这一特点,上述方法是可以非常有效地解决某些稀疏矩阵问题的。

8.1.9 关于结果精度的某些问题

现在讨论计算机上求得解答的精确度估计问题。提出这个问题是很自然的。因为实际问题中所提供的数据(系数矩阵和自由项的元素)或多或少总有一定误差(例如测量误差等等);将数据输入到计算机内并进行进制转换也会带来误差;有的问题中系数矩阵和自由项的元素是前面计算的结果,也不可避免地带有误差。由于这些原因,在计算机上实际求解的方程组的系数矩阵和自由项都包含有一定误差。这些误差对计算结果的精确度必定有影响。另外,计算机的字长是有限的,并且每作一次运算都要对结果进行舍入,在求解过程中这些舍入误差也会逐步积累,导致最终结果中有误差。所以有必要讨论一下计算结果的精

确度问题。

应该指出，要完全弄清楚这个问题不是一件容易的事情，我们仅对其中某些问题进行一些粗略的讨论。

(一)精确度的检验与病态矩阵

在计算机上求得方程组的近似解答之后，我们首先要问这一解答与方程组的精确解究竟相差多少？这个问题很难定量地回答。因为一般情况下方程组的精确解是不知道的，我们无法直接判定近似解与精确解有几位相同，只能通过一些间接的方式来估计解的精确度。

最简单的估计精确度的办法是把近似解 $\tilde{\boldsymbol{x}}$ 代入原来方程组(8.1.2)中去求出所谓"余量" $\boldsymbol{r}$:

$$\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$$

如果 $\boldsymbol{r}$ 的每个分量 r_i 都是小量或与自由项 $\boldsymbol{b}$ 的相应分量 b_i 相较都是小量(当 $b_i=0$ 时，r_i 本身应接近于零)，那么，一般就认为解是相当准确的，否则认为解是不准确的。

另外一个估计办法是任取一个已知向量(例如随机向量) $\boldsymbol{z}$, 计算出向量 $\boldsymbol{A}\cdot\boldsymbol{z}$(用较多的位数)，并以其为新的自由项，按同样方法再求解一次原来的方程组，即求解:

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{A}\cdot\boldsymbol{z}$$

所得解答记为 $\tilde{\boldsymbol{x}}$。如果计算过程没有舍入误差，$\tilde{\boldsymbol{x}}$ 应与 $\boldsymbol{z}$ 恒等。因求解过程中舍入误差的积累，$\tilde{\boldsymbol{x}}$ 与 $\boldsymbol{z}$ 之间将有差异。但由于求出我们所需要的近似解 $\tilde{\boldsymbol{x}}$ 与求出 $\tilde{\boldsymbol{x}}$ 的计算过程完全一样，系数矩阵亦相同，所以一般就认为两者的精确度是相同的。我们即可把 $\tilde{\boldsymbol{x}}$ 与 $\boldsymbol{z}$ 各分量之间相符合的最少位数作为近似解 $\tilde{\boldsymbol{x}}$ 的有效位数。

上述两种办法都是经常使用的。后者较前者有时更加可靠一些，一般能够得到关于近似解究竟有几位准确的一个数量概念。所以，是有其优点的。但其计算量较大(自然我们可以先算出 $\boldsymbol{A}\cdot\boldsymbol{z}$, 然后与 $\boldsymbol{b}$ 并列求解，以节省运算量)是一个很大的缺点。前者的优点则是，方法很简单、运算量少、对大多数实际问题也还是很可靠的。其缺点主要是从余量的大小无法断定近似解究竟有几位准确，所能得出的只是近似解准确与否的一个粗略概念。此外，这样的概念有时与实际情况可能还相差甚远。例如，我们考虑下列方程组:

$$\begin{bmatrix}0.2161, & 0.1441\\ 1.2969, & 0.8648\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}0.1440\\ 0.8642\end{bmatrix}$$

如果以 $\tilde{x}_1=0.991$, $\tilde{x}_2=-0.4870$ 代入这个方程组，将得出余量

$$\boldsymbol{r}=\begin{bmatrix}-0.00000001\\ 0.00000001\end{bmatrix}$$

很显然，若按照余量的大小来估计精确度，我们可以认为在取小数点后4位数字的情况下解答已足够精确了。然而，这一方程组的精确解却是 $x_1=2$, $x_2=-2$。这就说明，尽管余量已经很小，结果的精确度还可能很差。所以，上述按余量估计精确度的简单办法对于某些矩阵所相应的方程组来说是不可靠的。读者自然会问：对于什么样的矩阵会产生这种不可靠的现象？要回答这个问题，必须对矩阵的性质作进一步的讨论。

从上述例子可以看到，如果我们把方程组的自由项稍微变化一点，即变为:

$$\tilde{\boldsymbol{b}}=\begin{bmatrix}0.14400001\\ 0.86419999\end{bmatrix}$$

那么，准确解就由 $\boldsymbol{x}=(2,\ -2)^T$ 变为前述的 $\tilde{\boldsymbol{x}}$。因而，自由项的微小变化将引起解答的巨大

变化。这种现象通常叫做“病态”。同样地，方程组系数矩阵的微小变化有时亦会引起解答的巨大变化，这种现象也叫做“病态”。所以，更确切些说，如果系数矩阵或自由项的“微小”变化将引起方程组解答的巨大变化时，这个方程组就称为“病态”方程组，其系数矩阵就叫作对于解方程组(或求逆)来说的“病态矩阵”；反之，就称为“良态”方程组或者称“良态”矩阵。

应该指出，谈到“病态矩阵”的概念时，必须明确它是对什么而言的。因为对于解方程组(或求逆)来说是病态的矩阵，对于求特征值来说并不一定是病态的，反之亦然。所以，我们不能笼统地说某个矩阵是“病态”的。本章所说的“病态”则都是对于解方程组而言的。

还应指出，“病态”是系数矩阵本身的特性，与所用的计算工具和计算方法无关。但是，实际计算中“病态”的程度却是通过所用的计算工具等表现出来的。例如，计算机的字长愈长，“病态”现象在程度上就会相对地减轻。所以，我们前面在提及“微小”或“巨大”变化时均系相对而言，并无数量上的具体标准。一般来说，字长愈长，“微小”与巨大的相对范围就可愈大。

了解“病态”的概念以后，我们就可以说，前述按余量大小来判断近似解精确度的办法对于“病态”方程组来说一般是不可靠的。至于如何衡量一个矩阵是否病态的问题将在下面讨论。

(二)向量和矩阵的范数及其基本性质

如前所述，“病态”是矩阵本身的特性，故应该有一个只依赖于矩阵本身的衡量标准，这样才可能对不同矩阵的“病态”程度进行比较。

怎样比较不同矩阵对于解方程组来说的“病态”程度呢？为了解决这个问题，需要借助于向量和矩阵范数(或称模)的概念。这里，仅简单地叙述一下这些概念，详细内容可参阅[6]。

实向量 $\boldsymbol{x}=(x_1, x_2, \cdots, x_n)^T$ 的欧氏范数(或称欧氏模)是指如下实数：

$$\|\boldsymbol{x}\|=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}=\sqrt{\boldsymbol{x}^T\cdot\boldsymbol{x}} \tag{8.1.44}$$

与通常二维或三维空间的长度类似，向量的欧氏模具有如下性质：

(1) 任意非零向量 $\boldsymbol{x}$ 的欧氏模总是正实数，仅当 $\boldsymbol{x}$ 是零向量时，其范数才为零。

(2) 对于任何实数 c，下式均成立：

$$\|c\cdot\boldsymbol{x}\|=|c|\cdot\|\boldsymbol{x}\|$$

(3) 对于任意两个向量 $\boldsymbol{x}$, $\boldsymbol{y}$，有如下三角不等式：

$$\|\boldsymbol{x}+\boldsymbol{y}\|\leqslant\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$$

并且仅当 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 线性相关时上式中等号成立。

性质(1)、(2)的证明可直接从欧氏模的定义得到。性质(3)的证明复杂一些，其中要用到下列不等式：

$$|\boldsymbol{x}^T\cdot\boldsymbol{y}|\leqslant\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$$

其详细证明可参看[6]。

从向量的欧氏范数出发，我们可以定义相应的矩阵范数为[6]：

$$\|\boldsymbol{A}\|=\max_{\boldsymbol{x}\neq 0}\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}=\max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{A}\boldsymbol{x}\| \tag{8.1.45}$$

这样定义的矩阵欧氏范数就是矩阵 $\boldsymbol{A}^T\cdot\boldsymbol{A}$ 的最大特征值的平方根。当 $\boldsymbol{A}$ 为对称矩阵时，就是 $\boldsymbol{A}$ 的按模最大特征值。此外，它有如下性质：

(1) 对于任意非零矩阵 $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ 恒为正实数，仅当 $\boldsymbol{A}$ 为零矩阵时，其范数为零。

(2) 对于任意实数 c, 下式均成立:

$$\|c\cdot\boldsymbol{A}\|=|c|\cdot\|\boldsymbol{A}\|$$

(3) 若 $\boldsymbol{A}$、$\boldsymbol{B}$ 为同阶方阵，则有:

$$\|\boldsymbol{A}+\boldsymbol{B}\|\leqslant\|\boldsymbol{A}\|+\|\boldsymbol{B}\|$$

(4) 对于所有矩阵 $\boldsymbol{A}$ 及同维向量 $\boldsymbol{x}$, 下式均成立:

$$\|\boldsymbol{A}\cdot\boldsymbol{x}\|\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{x}\|$$

(5) 若 $\boldsymbol{A}$、$\boldsymbol{B}$ 为同阶方阵，则有:

$$\|\boldsymbol{A}\cdot\boldsymbol{B}\|\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{B}\|$$

前三个性质与向量范数的三个性质完全相同。所以，我们也可以把定义了范数的所有矩阵视为 n^2 维向量赋范空间。第四个性质也叫做“一致性”关系(或“相容性”关系)。凡是满足这一关系的向量范数与矩阵范数就叫做是“一致”的(或“相容”的)。一致性关系与性质(5)，在误差分析中是很有用处的。

上面定义的欧氏模与二维或三维空间中通常的长度概念是类似的，这就使我们可以在一般的 n 维向量空间中利用通常的几何直观。由于这个原因，欧氏模在误差分析中经常被利用。除欧氏模外，其他两种常用的向量范数为:

$$\|\boldsymbol{x}\|_I=\sum_{i=1}^{n}|x_i|$$
$$\|\boldsymbol{x}\|_\infty=\max_{1<i<n}|x_i| \tag{8.1.46}$$

按照(8.1.45)式，我们同样可以定义与上述向量范数相应的矩阵范数 $\|\boldsymbol{A}\|_I$ 和 $\|\boldsymbol{A}\|_\infty$:

$$\|\boldsymbol{A}\|_I=\max_{x\neq0}\frac{\|\boldsymbol{A}\boldsymbol{x}\|_I}{\|\boldsymbol{x}\|_I}$$
$$\|\boldsymbol{A}\|_\infty=\max_{x\neq0}\frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}$$

按照上述定义和(8.1.46)式，读者可以验证如下等式:

$$\|\boldsymbol{A}\|_I=\max_{1<j<n}\sum_{i=1}^{n}|a_{ij}|$$
$$\|\boldsymbol{A}\|_\infty=\max_{1<i<n}\sum_{j=1}^{n}|a_{ij}| \tag{8.1.47}$$

很容易验证前述的欧氏模的性质(1)～(5)对于这两种范数也是成立的。此外，不同的矩阵范数间还满足所谓范数等价定理: 若 $\|\boldsymbol{A}\|_\alpha$ 和 $\|\boldsymbol{A}\|_\beta$ 为任意两种矩阵范数，则必有如下关系式成立:

$$d\cdot\|A\|_\alpha\leqslant\|A\|_\beta\leqslant D\cdot\|A\|_\alpha$$

其中 d、D 为与 α、β 有关的常数。

由于篇幅所限，我们这里不再详细讨论这些内容，有兴趣的读者可以参阅[6]的第一章及[10、17]的有关章节。

上述三种向量范数可用如下公式统一地定义:

$$\|\boldsymbol{x}\|_p=\left(\sum_{j=1}^{n}|x_j|^p\right)^{1/p} \tag{8.1.48}$$

当 $p=2$ 时，就得出前面的欧氏范数，当 $p=1$ 或 ∞ 时，就分别得出(8.1.46)式中的两种范数。

由于这个原因，向量的欧氏模以及与其相应的矩阵的欧氏模有时也用符号 $\|\boldsymbol{x}\|_2$ 与 $\|\boldsymbol{A}\|_2$ 表示，并称为 2-模，而 $\|\boldsymbol{A}\|_1$ 与 $\|\boldsymbol{A}\|_\infty$ 则分别称之为 1-模及 ∞-模。

最后，我们还要指出 1-模和 ∞-模在运用几何直观方面不如欧氏模那样简单。例如，正交变换下向量和矩阵的欧氏模不变，而这两种模则可能发生变化。不过，由于它们较欧氏模容易计算，许多场合还是很有用处的。

（三）矩阵的条件数

现在再来讨论如何比较不同矩阵病态程度的问题。为此，需要弄清系数矩阵和自由项有一个微小的变化时，方程组的解是怎样变化的。这个问题也叫做“摄动分析”。如果用 $\Delta\boldsymbol{A}$、$\Delta\boldsymbol{b}$ 表示方程组(8.1.2)的系数矩阵 $\boldsymbol{A}$ 和自由项 $\boldsymbol{b}$ 的微小变化，解 $\boldsymbol{x}$ 的相应变化记为 $\Delta\boldsymbol{x}$，自然就有：

$$(\boldsymbol{A}+\Delta\boldsymbol{A})\cdot(\boldsymbol{x}+\Delta\boldsymbol{x})=\boldsymbol{b}+\Delta\boldsymbol{b} \tag{8.1.49}$$

将上式展开，并利用(8.1.2)，便得到：

$$\boldsymbol{A}\cdot\Delta\boldsymbol{x}+\Delta\boldsymbol{A}\cdot\boldsymbol{x}+\Delta\boldsymbol{A}\cdot\Delta\boldsymbol{x}=\Delta\boldsymbol{b}$$

对这一等式两端左乘以矩阵 $\boldsymbol{A}^{-1}$，并进行移项，便得到：

$$\Delta\boldsymbol{x}=-\boldsymbol{A}^{-1}\cdot\Delta\boldsymbol{A}\cdot\boldsymbol{x}-\boldsymbol{A}^{-1}\cdot\Delta\boldsymbol{A}\cdot\Delta\boldsymbol{x}+\boldsymbol{A}^{-1}\cdot\Delta\boldsymbol{b}$$

将上式两端取范数（这里我们对范数符号不加下标，以表示任取一种范数），并利用向量和相应矩阵范数的一致性关系，即得：

$$\|\Delta\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\|\cdot\|\boldsymbol{x}\|+\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\|\cdot\|\Delta\boldsymbol{x}\|+\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{b}\|$$

再将其两端同除以 $\|\boldsymbol{x}\|$，并移项得：

$$(1-\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\|)\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{A}^{-1}\|\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{A}\|\|\boldsymbol{x}\|}+\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\| \tag{8.1.50}$$

另外，对(8.1.2)式两端取范数可得：

$$\|\boldsymbol{b}\|\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{x}\|$$

这样，我们便可将(8.1.50)中的 $\|\boldsymbol{A}\|\cdot\|\boldsymbol{x}\|$ 换成 $\|\boldsymbol{b}\|$，从而得到：

$$(1-\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\|)\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{A}^{-1}\|\left(\frac{\|\Delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}+\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right) \tag{8.1.51}$$

如果矩阵的微小变化 $\Delta\boldsymbol{A}$ 满足如下条件：

$$\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\|\leqslant\alpha<1 \tag{8.1.52}$$

则从(8.1.51)两端除以 $1-\|\boldsymbol{A}^{-1}\|\cdot\|\Delta\boldsymbol{A}\|$，并令：

$$P(\boldsymbol{A})=\|\boldsymbol{A}^{-1}\|\cdot\|\boldsymbol{A}\| \tag{8.1.53}$$

便得到：

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{P(\boldsymbol{A})}{1-P(\boldsymbol{A})\dfrac{\|\Delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}}\cdot\left(\frac{\|\Delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}+\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right) \tag{8.1.54}$$

这一不等式说明了方程组解答的相对变化与系数矩阵和自由项的相对变化间的关系。从其中可以看出，只要满足条件(8.1.52)，对于系数矩阵或自由项的同样的相对变化来说，$P(\boldsymbol{A})$ 愈大，解的相对变化就可能愈大，$P(\boldsymbol{A})$ 愈小，解的相对变化就应愈小。所以，$P(\boldsymbol{A})$ 在某种程度上刻划了方程组的解答对于原始数据变化的灵敏程度，也就是刻划了方程组“病态”的程度。此外，尽管(8.1.54)为一不等式，在很多情况下是“过估”的，但容易找到使其中等号接近成立的矩阵。所以，很自然地应把 $P(\boldsymbol{A})$ 看作矩阵对于解方程组来说“病态”程度

的一种度量。通常把 $P(\boldsymbol{A})$ 称为矩阵 $\boldsymbol{A}$(对于所取范数)的条件数(或性态数)。$P(\boldsymbol{A})$愈大,矩阵 $\boldsymbol{A}$ 对于解方程组(或求逆)来说就愈病态。与谈到"病态"的定义时类似,究竟条件数多大矩阵才算病态,一般来说是没有具体标准的,也只是相对而言。

从条件数 $P(\boldsymbol{A})$的定义易知,如果 $\boldsymbol{A}$ 是实对称矩阵并取欧氏模,那么便有:

$$\|\boldsymbol{A}\|_2=\sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})}=|\lambda_1|$$

$$\|\boldsymbol{A}^{-1}\|_2=\sqrt{\lambda_{\max}(\boldsymbol{A}^{-T}\boldsymbol{A}^{-1})}=\sqrt{\frac{1}{\lambda_{\min}(\boldsymbol{A}^T\cdot\boldsymbol{A})}}=\frac{1}{|\lambda_n|}$$

所以:

$$P(\boldsymbol{A})=|\lambda_1|/|\lambda_n| \tag{8.1.55}$$

其中 λ_1 与 λ_n 分别为矩阵 $\boldsymbol{A}$ 的按模最大和最小特征值。实对称矩阵条件数的表达式(8.1.55)以后还会常常碰到。

(四)舍入误差问题

前面已经讨论了方程组原始数据变化对其解答的影响,并引入了条件数概念。下面简单叙述一下计算过程中的舍入误差对求得解答的影响。仔细分析舍入误差积累问题是比较复杂和麻烦的。同时,目前所得到的理论结果也与实际出入较大。所以,这里不去涉及误差分析的细节,只引用一些必要的结论。

现在经常采用的一种严格误差分析方法是所谓"向后误差分析法"。其基本思想是把计算过程中的舍入误差的影响归结为原始数据变化的影响。也就是说,去找出原始数据的某种变化 $\Delta\boldsymbol{A}$、$\Delta\boldsymbol{b}$,使其对最终解答的影响同求解过程中舍入误差的影响等效。这样计算所得的近似解 $\tilde{\boldsymbol{x}}$ 就将严格地满足如下方程式:

$$(\boldsymbol{A}+\Delta\boldsymbol{A})\tilde{\boldsymbol{x}}=\boldsymbol{b}+\Delta\boldsymbol{b}$$

找到这样的 $\Delta\boldsymbol{A}$、$\Delta\boldsymbol{b}$ 以后,解答精确度的估计问题就化为前面讨论过的"摄动分析"问题了。

显然,对于不同的计算方法,其等效的原始数据变化 $\Delta\boldsymbol{A}$、$\Delta\boldsymbol{b}$ 是不同的。此外,从(8.1.54)式可知,要得出解答的精确度估计,只需知道$\|\Delta\boldsymbol{A}\|$ 和 $\|\Delta\boldsymbol{b}\|$即已足够。对于本节所讨论的各种直接法已进行过详细的误差分析(见[16、17]),得出了相应的 $\|\Delta\boldsymbol{A}\|$ 和 $\|\Delta\boldsymbol{b}\|$ 的上界。例如,对于高斯消去法,其结果为:

$$\begin{aligned}&\|\Delta\boldsymbol{A}\|_\infty\leqslant c_1\cdot G\cdot a\cdot(2n^2+n^3)\cdot 2^{-t}\\&\|\Delta\boldsymbol{b}\|_\infty=0\end{aligned} \tag{8.1.56}$$

其中,a 为矩阵 $\boldsymbol{A}$ 的按模最大元素,即

$$a=\max_{1\leqslant i,j\leqslant n}|a_{ij}|$$

G 为消去过程中矩阵元素的最大增长因子,即

$$G=\max_{\substack{1\leqslant i,j\leqslant n\\1\leqslant k\leqslant n}}|a_{ij}^{(k)}|/a$$

c_1 为近于1的某个常数;t 为所用计算机的字长(指尾数部分而言)。

直接分解法的结果完全与消去法类似。

对于平方根法,由于矩阵 $\boldsymbol{L}$ 的任一元素 l_{ij} 之模均不超过 $a=\max\limits_{1\leqslant i,j\leqslant n}|a_{i,j}|$,故其结果为:

$$\begin{aligned}&\|\Delta\boldsymbol{A}\|_\infty\leqslant c_2\cdot a\cdot(2n^2+n^3)\cdot 2^{-t}\\&\|\Delta\boldsymbol{b}\|_\infty=0\end{aligned} \tag{8.1.57}$$

对于镜像映射法,其结果为:

$$\|\Delta A\|_2 \leqslant c_3 \cdot n^2 \cdot 2^{-t} \cdot \|A\|_2$$
$$\|\Delta b\|_2 \leqslant c_4 \cdot n^2 \cdot 2^{-t} \cdot \|b\|_2 \tag{8.1.58}$$

其中 c_3、c_4 均为接近于1的常数。

将上面这些结果代入(8.1.54)，并利用(8.1.52)，立即可得近似解 $\tilde{x}$ 的精确度估计如下：

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \leqslant \begin{cases} P(A)\left(\frac{c_1}{1-\alpha} \cdot G\right)(2n^2+n^3) \cdot 2^{-t} & \text{（消去法或直接分解法）} \\ P(A)\left(\frac{c_2}{1-\alpha}\right)(2n^2+n^3) \cdot 2^{-t} & \text{（平方根法）} \end{cases} \tag{8.1.59}$$

$$\frac{\|x-\tilde{x}\|_2}{\|x\|_2} \leqslant N(A)\left(\frac{c_3+c_4}{1-\alpha}\right) \cdot n^2 \cdot 2^{-t} \quad \text{（镜像映射法）}$$

其中 $P(A)=\|A\|_\infty \cdot \|A^{-1}\|_\infty$, $N(A)=\|A\|_2 \cdot \|A^{-1}\|_2$。

从这些估计式可以看出无论采用哪一种解法，条件数愈大，舍入误差的影响愈严重。因而，条件数的大小在一定程度上表征了求解该方程组过程中舍入误差影响的大小。所以，条件数对于刻划方程组的性质来说是十分重要的。此外，从这些估计式中也可看到，方程组的阶数愈高，计算机的字长愈短，舍入误差的影响也就愈大。

应该指出，这些估计式均是严格的上界，往往是很保守的。虽然从中可以粗略看到各种因素在舍入误差积累中的作用，但用来作为精确度的估计还很不理想。例如，经验证明对于消去法将有如下结果：

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \leqslant c \cdot G \cdot n \cdot 2^{-t} \cdot P(A) \tag{8.1.60}$$

这显然比(8.1.59)中的上界小得多。此外，矩阵的条件数往往事先并不知道，用(8.1.59)来作近似解的精确度估计也是不大实用的。目前大多数严格的误差估计式均有这个缺点。也有人用概率论的方法来估计舍入误差的影响，所得的结果虽比上述严格的上界估计好一些，但实际使用价值也不很大。总之，关于求解过程中舍入误差的分析问题，目前还没有得到完满的解决。

（五）提高精确度的某些措施

从前面讨论读者可以看到，关于结果精确度的问题目前在理论上还没有得到完满的解决。实际计算中往往不去使用误差分析的理论结果，经常采取的检验办法则是从问题的物理背景出发，去分析解答的规律与趋势，看其是否与客观规律相符合；并且也同时采取某些简单措施，例如，代入方程去验算（即求余量 r），或用已经知道准确解答的自由项再解一次，或作某些后验估计（参见下面的讨论）等等。这样从各个方面综合进行判断，一般来说还是可以得到比较满意的结论的。

如果检验过程中发现解的精确度较差，不能满足需要，我们就应该采取一些措施来改进解的精确度。为此，首先要搞清楚导致结果不准确的真正原因，以便对症下药。通常有两个方面的原因：一是问题本身病态，因而计算结果不准确；另一个是问题本身并非很病态，主要是计算方法选择得不合理（例如，对一般矩阵使用了顺序消去法等等）。后一种情况，只要我们在选择方法时多加注意，一般是可以避免的。主要问题在于如何处理前一种情况。

由于矩阵的病态度量（例如条件数）事先难以知道，所以实际计算中只能通过一些现象间接地进行判断。一般来说，遇到下列几种情况，对于解方程组（或求逆）来说，矩阵就有可

能是病态的:

(1) 矩阵元素间的数量级相差很远,大的很大,小的很小,并且无一定规则等等。

(2) 矩阵的行列式值相对来说很小,或其某些行(或者列)近似地线性相关。

(3) 消去过程中出现严重的有效位消失或数值很小的主元素, 以及由数量级相差很远的两行进行加减等等。

(4) 余量已很小,但解仍不符合规律。

这时, 我们就可以按"病态"情形来加以处理(当然, 我们可将矩阵元素或自由项改变一个小的数值,再进行一次求解。如果解的变化很大, 那就可以肯定矩阵是病态的)。 通常的处理办法有下列几种:

(1) 对矩阵的行(列)适当地引入比例因子(为了减少舍入误差,比例因子通常取为 2 的幂次),使矩阵各行(或列)的元素均在 ± 1 之间, 且各行(列)之模大致相等。 这种引入比例因子(也叫作行列平衡)的办法有时能够提高一些精确度,有时也可能使精确度变得更差,所以不能随意使用。对于消去法来说,引入比例因子主要是改变消去时所使用的主元素位置。如果引入比例因子后仍然按以前位置的主元素进行消去, 解的精度将不会有什么变化。 除了上述引入比例因子的办法以外,还有一些"平衡"矩阵的办法, 它们大都比较繁琐, 效果也并不显著,目前还没有一个既方便而又有效的通用办法。实践经验表明,大多数情况下对行和列同时引入比例因子会改善结果的精确度,但其理论根据尚不完全清楚。

(2) 采取双倍字长(或多倍字长)进行运算。这是一个比较有效的措施。由于字长增加一倍(或几倍),前述的各种病态现象,在程度上一般都会有很大的减轻与改善。其缺点主要是计算时间将大为增加(一般要增至单字长运算的几倍至十几倍),存储量也将增加两倍(或几倍)左右。值得推荐的是采用所谓双倍内积的办法,即在作 $\sum_i a_i \cdot b_i$ 型计算(例如求三角形分解式的各元素)时,每一乘积 $a_i \cdot b_i$ 均按双倍位长的数与其它项累加, 最终再舍入成单字长数。这样就保证了关键计算步骤的精确度, 最终结果精度将显著提高。 同时也比普遍地使用双倍字长运算的办法节省很多计算时间。

(3) 迭代改进的办法

这是目前最成功的改进解答精确度的办法之一。 只要矩阵不是非常病态, 通过这一方法总可以得到相应方程组精确解的较好近似值(参见[8、11])。其计算步骤大体如下:

(i) 用单字长运算(或双倍内积运算)将矩阵分解为下三角矩阵 $\boldsymbol{L}$ 与上三角矩阵 $\boldsymbol{U}$ 的乘积,并将 $\boldsymbol{L}$、$\boldsymbol{U}$ 的元素(单字长数)存放起来(这一步应采取列主元素法或全主元素法)。

(ii) 用单字长运算解方程组:

$$\begin{cases} \boldsymbol{L}\boldsymbol{y}_0 = \boldsymbol{b} \\ \boldsymbol{U}\boldsymbol{x}_0 = \boldsymbol{y}_0 \end{cases}$$

得出零次近似值 $\boldsymbol{x}_0$。

(iii) 用双倍内积运算计算余量 $\boldsymbol{r}_0$, 然后将 $\boldsymbol{r}_0$ 舍入成单字长数:

$$\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$$

(iv) 用单字长运算解方程组:

$$\begin{cases} \boldsymbol{L}\boldsymbol{y}_1 = \boldsymbol{r}_0 \\ \boldsymbol{U}\tilde{\boldsymbol{x}}_1 = \boldsymbol{y}_1 \end{cases}$$

求出解的修正量 $\tilde{\boldsymbol{x}}_1$。

(v) 求出修正后的解 $\boldsymbol{x}_1 = \boldsymbol{x}_0 + \tilde{\boldsymbol{x}}_1$。(注意! $[\|\tilde{\boldsymbol{x}}\|/\|\boldsymbol{x}_1\|]$ 就是解答 $\boldsymbol{x}_0$ 的有效位数,有时

这是有用处的)。

(vi) 将 $\boldsymbol{x}_1$ 代替 $\boldsymbol{x}_0$, 重复 (iii)、(iv)、(v)。

这样,我们得出一个近似解的序列:

$$\boldsymbol{x}_0,\ \boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_k,\ \cdots$$

如果 k 大于某个 k_0 以后，所有 $\boldsymbol{x}_k$ 都近似地等于某个单字长向量 $\boldsymbol{x}^*$，那么除极个别特殊情况外，$\boldsymbol{x}^*$ 即为真解 $\boldsymbol{A}^{-1}\cdot\boldsymbol{b}$ 的正确舍入值。

上述迭代改进过程只将矩阵 $\boldsymbol{A}$ 分解一次(其计算量为 $n^3/3$)。其后每迭代一次只需解两个三角形方程组(其计算量为 n^2), 并且, 一般情况下(即矩阵不是十分病态), 迭代次数只有 4~5 次。因而,整个工作量比起只解一次方程组来说增加不多。

还应指出,计算各次余量 $\boldsymbol{r}_k$ 时,必须用双倍内积运算,否则 $\boldsymbol{x}_k$ 的精度得不到改进,甚至序列$\{\boldsymbol{x}_k\}$不收敛。另外,当矩阵十分病态时(例如$\|\boldsymbol{A}\|\cdot\|\boldsymbol{A}^{-1}\|>2^t$),序列$\{\boldsymbol{x}_k\}$也不收敛。这时必须进一步采取其他措施(例如采用多字长运算重新求解)。

(4) 有许多工程实际问题, 按其物理背景来说,相应的方程组不应该是病态的。然而,由于问题的提法或处理不妥当, 有时会使方程组变成病态的。所以, 在我们处理病态问题时, 也应从物理背景方面认真检查“病态”产生的原因, 以便发现问题形成过程中的不妥之处,并加以纠正。有时,这种办法可以有效地解决问题。

§ 8.2 解线性代数方程组的迭代法

8.2.1 前　　言

(一)§ 8.1 中所讨论的直接法对于阶数不是很高的问题是非常有效的,这种场合一般不使用本节所讨论的迭代法。然而,对于阶数很高的稀疏矩阵,尽管提出了很多特殊的直接法来处理它们,在运算量和存储量的节省方面也取得了很大的进展,但仍然难于完全克服存储需要量大的缺点,有时常常需要采用本节所讨论的迭代法来解决问题。

迭代法由于不需要存储系数矩阵的零元素(有时采取由程序临时产生非零元素的办法,系数矩阵也无需存储),所以占用的存储单元少。同时,程序也较简单,对于许多问题(例如,二阶椭圆型方程边值问题的差分方程组求解问题等), 收敛较快。因而, 有时能够很有效地解决一些高阶问题。但是, 对于某些问题迭代法可能发散或收敛很慢, 以致失去使用价值。这种情况下,仍以采用直接法为宜。

迭代法的基本思想是去构成一个向量序列$\{\boldsymbol{u}^{(k)}\}$,使其收敛至某个极限向量 $\boldsymbol{u}^*$,并且$\boldsymbol{u}^*$就是要求解的方程组:

$$\boldsymbol{A}\boldsymbol{u}=\boldsymbol{b} \tag{8.2.1}$$

的准确解。

我们先以三阶方程组为例来说明迭代法的计算过程。假定要求解的方程组(8.2.1)是如下三阶方程组:

$$\begin{cases} a_{11}u_1+a_{12}u_2+a_{13}u_3=b_1 \\ a_{21}u_1+a_{22}u_2+a_{23}u_3=b_2 \\ a_{31}u_1+a_{32}u_2+a_{33}u_3=b_3 \end{cases} \tag{8.2.2}$$

不妨假设 $a_{ii}\neq 0(i=1,\ 2,\ 3)$, 于是,我们用 a_{11} 去除第一个方程式两端,并将其改写为:

$$u_1 = -\frac{a_{12}}{a_{11}}u_2 - \frac{a_{13}}{a_{11}}u_3 + \frac{b_1}{a_{11}}$$

对第二、三两个方程也作类似处理，并令：

$$\begin{cases} b_{ij} = -a_{ij}/a_{ii} \quad (i \neq j) \\ c_i = b_i/a_{ii} \end{cases} \tag{8.2.3}$$

就可以把(8.2.2)化为下列等价(即有相同解答)形式：

$$\begin{cases} u_1 = \qquad\quad b_{12}u_2 + b_{13}u_3 + c_1 \\ u_2 = b_{21}u_1 \qquad\quad + b_{23}u_3 + c_2 \\ u_3 = b_{31}u_1 + b_{32}u_2 \qquad\quad + c_3 \end{cases} \tag{8.2.4}$$

任选一组数 $u_1^{(0)}$、$u_2^{(0)}$、$u_3^{(0)}$ 作为方程组的近似解(通常称为初值，或零次近似解)，将其代入(8.2.4)右端即可求出一组新的数值 $u_1^{(1)}$、$u_2^{(1)}$、$u_3^{(1)}$。即是说：

$$\begin{cases} u_1^{(1)} = \qquad\quad b_{12}u_2^{(0)} + b_{13}u_3^{(0)} + c_1 \\ u_2^{(1)} = b_{21}u_1^{(0)} \qquad\quad + b_{23}u_3^{(0)} + c_2 \\ u_3^{(1)} = b_{31}u_1^{(0)} + b_{32}u_2^{(0)} \qquad\quad + c_3 \end{cases}$$

把这组新值作为改进后的近似解(或称一次近似解)，再将其代入(8.2.4)右端即可求得二次近似解，并如此反复迭代下去，就得到一个近似解序列。其一般的迭代计算公式即可写为：

$$\begin{cases} u_1^{(k+1)} = \qquad\quad b_{12}u_2^{(k)} + b_{13}u_3^{(k)} + c_1 \\ u_2^{(k+1)} = b_{21}u_1^{(k)} \qquad\quad + b_{23}u_3^{(k)} + c_2 \\ u_3^{(k+1)} = b_{31}u_1^{(k)} + b_{32}u_2^{(k)} \qquad\quad + c_3 \end{cases} \tag{8.2.5}$$

或者

$$\begin{cases} u_1^{(k+1)} = a_{11}^{-1}\cdot a_{12}u_2^{(k)} + a_{11}^{-1}\cdot a_{13}u_3^{(k)} + a_{11}^{-1}\cdot b_1 \\ u_2^{(k+1)} = a_{22}^{-1}\cdot a_{21}u_1^{(k)} + a_{22}^{-1}\cdot a_{23}u_3^{(k)} + a_{22}^{-1}\cdot b_2 \\ u_3^{(k+1)} = a_{33}^{-1}\cdot a_{31}u_1^{(k)} + a_{33}^{-1}\cdot a_{32}u_2^{(k)} + a_{33}^{-1}\cdot b_3 \end{cases}$$

$$(k=0,\ 1,\ 2,\ \cdots)$$

如果采用下列矩阵符号：

$$\boldsymbol{D} = \begin{pmatrix} a_{11} & & \boldsymbol{0} \\ & a_{22} & \\ \boldsymbol{0} & & a_{33} \end{pmatrix},\quad \boldsymbol{B} = \begin{pmatrix} 0 & b_{12} & b_{13} \\ b_{21} & 0 & b_{23} \\ b_{31} & b_{32} & 0 \end{pmatrix},\quad \boldsymbol{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix},$$

$$\boldsymbol{u}^{(k)} = \begin{pmatrix} u_1^{(k)} \\ u_2^{(k)} \\ u_3^{(k)} \end{pmatrix},\quad (k=0,\ 1,\ 2,\ \cdots)$$

显然有：

$$\boldsymbol{B} = \boldsymbol{D}^{-1}\cdot(\boldsymbol{D}-\boldsymbol{A}),\quad \boldsymbol{c} = \boldsymbol{D}^{-1}\cdot\boldsymbol{b}$$

迭代公式(8.2.5)即可记为：

$$\begin{cases} \boldsymbol{u}^{(0)},\ \text{任取初始值} \\ \boldsymbol{u}^{(k+1)} = \boldsymbol{B}\boldsymbol{u}^{(k)} + \boldsymbol{c},\quad (k=0,\ 1,\ 2,\ \cdots) \end{cases} \tag{8.2.6}$$

按(8.2.5)计算下去，当 $k\to\infty$ 时，若 $u_1^{(k)}$、$u_2^{(k)}$、$u_3^{(k)}$ 分别收敛于某个极限值 u_1^*、u_2^*、u_3^*，则对(8.2.5)两端取极限即可看出：

$$\begin{cases} u_1^* = \qquad\quad b_{12}u_2^* + b_{13}u_3^* + c_1 \\ u_2^* = b_{21}u_1^* \qquad\quad + b_{23}u_3^* + c_2 \\ u_3^* = b_{31}u_1^* + b_{32}u_2^* \qquad\quad + c_3 \end{cases}$$

所以，u_1^*、u_2^*、u_3^* 就是(8.2.4)的解，因而，也是原方程组(8.2.2)的解。

由于向量$\boldsymbol{u}^{(k)}$的各分量均有极限，各分量的极限值u_i^*亦组成一个向量$\boldsymbol{u}^*=(u_1^*, u_2^*, u_3^*)^T$，自然，此时我们就说向量序列$\{\boldsymbol{u}^{(k)}\}$有极限$\boldsymbol{u}^*$。这样一来，按迭代公式(8.2.5)或(8.2.6)，我们就得到一个向量序列$\{\boldsymbol{u}^{(k)}\}$，它收敛至极限向量$\boldsymbol{u}^*$，并且$\boldsymbol{u}^*$就是方程组(8.2.2)的准确解。以上就是通常所谓简单迭代法的计算过程。任一迭代法所构造的向量序列$\{\boldsymbol{u}^{(k)}\}$收敛至方程组的准确解$\boldsymbol{u}^*$时，就称为该迭代法收敛。否则叫做该迭代法不收敛或发散。

从上述例子可以看出，迭代法实际上就是一种从已有近似解计算新的近似解的规则。计算规则不同，迭代法也就不同。(8.2.6)中新的近似解$\boldsymbol{u}^{(k+1)}$是已有近似解的线性函数，这种迭代法叫做线性迭代法，还有一些迭代法从已有近似解计算新的近似解所用的公式是非线性的，就叫做非线性迭代法。此外，(8.2.6)中$\boldsymbol{u}^{(k+1)}$仅仅是其前面一次近似解$\boldsymbol{u}^{(k)}$的函数，这种迭代法叫做一阶迭代法。如果不仅是前一次近似解，还是其前面相邻的p个近似解$\boldsymbol{u}^{(k)}, \boldsymbol{u}^{(k-1)}, \cdots, \boldsymbol{u}^{(k-p+1)}$的函数，则叫作$p$阶的迭代法。最后，还可看到(8.2.6)中无论$k$为何值，计算$\boldsymbol{u}^{(k+1)}$的规则均是不变的，这种迭代法称之为定常迭代法。如果计算新的近似解的规则是随迭代过程而变的，就叫作非定常迭代法。按照上述分类方式，显然，我们可以说简单迭代法是一种一阶线性定常迭代法。

一阶线性定常迭代法的一般形式可以写为：

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{G}\boldsymbol{u}^{(k)} + \boldsymbol{d} \quad (k=0, 1, 2, \cdots) \tag{8.2.7}$$

其中$\boldsymbol{G}$为n阶方阵，通常称之为该迭代法的迭代矩阵，$\boldsymbol{d}$为n维常向量，$\boldsymbol{G}$和$\boldsymbol{d}$均与k无关。除前述的简单迭代法外，高斯-赛德尔(Gauss-Seidel)迭代法和逐次松弛法，亦属于此范围。我们将首先在8.2.2节中讨论这一类方法，特别要较详细讨论逐次松弛法。除此而外，某些非定常的线性迭代法，例如切比雪夫半迭代法，有时也是有用的，我们放在8.2.3节中作一简单叙述。8.2.4节是讨论分块迭代法。8.2.5节中讨论一种常用的非线性迭代法——共轭斜量法。

应该指出，迭代法的使用与问题特点有很密切的联系。各种实际问题中使用迭代法的经验以及针对问题特点所作的详细讨论等，是很重要的。特别是迭代法在椭圆型方程边值问题数值求解中的应用更是如此。关于这方面的问题读者可参阅本书第十三章。这里，只是给出上述几种迭代法的一些简单概念和常用结果，其它问题不可能在此作详细叙述，这一点请读者注意。

(二)讨论迭代法的过程中，经常要遇到矩阵或向量序列的极限概念，我们先来简单叙述一下这些概念。

定义 2.1　一个向量序列$\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \boldsymbol{v}^{(3)}, \cdots, \boldsymbol{v}^{(k)}, \cdots$(以后简记$\{\boldsymbol{v}^{(k)}\}$)叫作收敛的，是指由每个向量的相应分量所组成的$n$个数序列$\{v_i^{(k)}\}$ $(i=1, 2, \cdots, n)$是收敛的。以这n个数序列的极限v_i为相应分量的向量$\boldsymbol{v}$，就称为向量序列$\{\boldsymbol{v}^{(k)}\}$的极限。即是说：

$$\lim_{k\to\infty} \boldsymbol{v}^{(k)} = \boldsymbol{v} \Leftrightarrow \lim_{k\to\infty} v_i^{(k)} = v_i \quad (i=1, 2, \cdots, n)$$

其中$v_i^{(k)}$和v_i分别为向量$\boldsymbol{v}^{(k)}$和$\boldsymbol{v}$的第i个分量。

完全类似，我们有如下定义：

定义 2.2　一个矩阵序列 $\boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \boldsymbol{A}^{(3)}, \cdots, \boldsymbol{A}^{(k)}, \cdots$（亦简记为 $\{\boldsymbol{A}^{(k)}\}$）收敛于矩阵 $\boldsymbol{A}$，是指如下 n^2 个数量的极限关系式成立：

$$\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij} \quad (i, j = 1, 2, \cdots, n)$$

其中 $a_{ij}^{(k)}$ 及 a_{ij} 分别为矩阵 $\boldsymbol{A}^{(k)}$ 及 $\boldsymbol{A}$ 的 i 行 j 列元素。

从上述定义和本章 § 8.1 中关于矩阵及向量范数的定义，我们可以直接验证下面两个定理的正确性。

定理 2.1　向量序列 $\{\boldsymbol{v}^{(k)}\}$ 收敛于向量 $\boldsymbol{v}$ 的充要条件是对于任一种向量范数均有：

$$\lim_{k\to\infty} \|\boldsymbol{v}^{(k)} - \boldsymbol{v}\| = 0$$

定理 2.2　矩阵序列 $\{\boldsymbol{A}^{(k)}\}$ 收敛于矩阵 $\boldsymbol{A}$ 的充要条件是对于任一种矩阵范数均有：

$$\lim_{k\to\infty} \|\boldsymbol{A}^{(k)} - \boldsymbol{A}\| = 0$$

定理的证明很简单。若 $\{\boldsymbol{v}^{(k)}\}$ 收敛于向量 $\boldsymbol{v}$，根据定义 2.1，显然有 $\boldsymbol{v}^{(k)} - \boldsymbol{v}$ 收敛于零向量，自然其任何一种范数均收敛于零。反之，若 $\boldsymbol{v}^{(k)} - \boldsymbol{v}$ 的任一种范数收敛于零，则因只有零向量的范数才会为零，所以 $\boldsymbol{v}^{(k)} - \boldsymbol{v}$ 必定收敛于零向量，即 $\lim\limits_{k\to\infty} \boldsymbol{v}^{(k)} = \boldsymbol{v}$。这样就证明了定理 2.1。完全类似地可以证明定理 2.2。

以后的讨论中常常要用到矩阵的谱半径概念，其定义如下：

定义 2.3　任意方阵 $\boldsymbol{G}$ 的按模最大特征值的模称为方阵 $\boldsymbol{G}$ 的谱半径，并以符号 $S(\boldsymbol{G})$ 表示。即是说：

$$S(\boldsymbol{G}) = \max_{1 \leqslant i \leqslant n} |\mu_i| \tag{8.2.8}$$

其中 μ_i 为矩阵 $\boldsymbol{G}$ 的特征值。

下面我们来证明一个关于矩阵的幂次收敛于零的充要条件，对于讨论迭代法的收敛性问题，这个充要条件是很重要的。

定理 2.3　任意方阵 $\boldsymbol{G}$ 的 k 次乘幂 $\boldsymbol{G}^k$，当 $k\to\infty$ 时收敛于零的充要条件是方阵 $\boldsymbol{G}$ 的谱半径小于 1。

证明：假设 μ 是 $\boldsymbol{G}$ 的按模最大特征值，其相应特征向量为 $\boldsymbol{x}$，对等式 $\mu\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x}$ 两边任取一种范数，则从范数的基本性质可以推得：

$$|\mu| \cdot \|\boldsymbol{x}\| \leqslant \|\boldsymbol{G}\| \cdot \|\boldsymbol{x}\| \quad 或者 \quad S(\boldsymbol{G}) = |\mu| \leqslant \|\boldsymbol{G}\|$$

即是说，矩阵 $\boldsymbol{G}$ 的谱半径应不超过其任何一种范数。

若 $\boldsymbol{G}^k \to 0\ (k\to\infty)$，但有 $S(\boldsymbol{G}) \geqslant 1$，则我们可以推知 $S(\boldsymbol{G}^k) = [S(\boldsymbol{G})]^k \geqslant 1$，对于所有 k 皆成立。于是我们便有：

$$\|\boldsymbol{G}^k\| \geqslant S(\boldsymbol{G}^k) \geqslant 1 \quad (对于所有的\ k)$$

根据定理 2.2，这与 $\boldsymbol{G}^k \to 0$ 是矛盾的，所以，若 $\boldsymbol{G}^k \to 0$，应有 $S(\boldsymbol{G}) < 1$。必要性即得证明。

为了证明充分性，我们需要利用矩阵的若当标准型的结果（参见[4]第六章）。即是说对于方阵 $\boldsymbol{G}$，可以找到某个非奇矩阵 $\boldsymbol{P}$，使得：

$$\boldsymbol{P}^{-1}\boldsymbol{G}\boldsymbol{P} = \boldsymbol{J} = \begin{pmatrix} \boldsymbol{J}_1 & & \boldsymbol{0} \\ & \boldsymbol{J}_2 & \\ & & \ddots \\ \boldsymbol{0} & & & \boldsymbol{J}_s \end{pmatrix} \tag{8.2.9}$$

其中 $\boldsymbol{J}_i$ 为下列形状的 m_i 阶若当块：

$$\boldsymbol{J}_i=\begin{pmatrix} \mu_i & 1 & & \boldsymbol{0} \\ & \mu_i & 1 & \\ & & \ddots & 1 \\ \boldsymbol{0} & & & \mu_i \end{pmatrix} \quad (i=1,\ 2,\ \cdots,\ s)$$

这里 μ_i 为矩阵 $\boldsymbol{G}$ 的特征值，并且 $m_1+m_2+\cdots+m_s=n$。矩阵 $\boldsymbol{J}$ 就称为矩阵 $\boldsymbol{G}$ 的若当标准型。

从(8.2.9)容易推知：

$$\boldsymbol{G}^k=\boldsymbol{P}\boldsymbol{J}^k\boldsymbol{P}^{-1}=\boldsymbol{P}\cdot\begin{bmatrix} \boldsymbol{J}_1^k & & \boldsymbol{0} \\ & \boldsymbol{J}_2^k & \\ & & \ddots \\ \boldsymbol{0} & & \boldsymbol{J}_s^k \end{bmatrix}\cdot\boldsymbol{P}^{-1} \tag{8.2.10}$$

根据 $\boldsymbol{J}_i$ 的特殊形状，经过简单计算即可得知：

$$\boldsymbol{J}_i^k=\begin{bmatrix} \mu_i^k, & k\mu_i^{k-1}, & C_k^2\mu_i^{k-2}, & \cdots & C_k^{m_i-1}\mu_i^{k-m_i+1} \\ & \mu_i^k, & k\mu_i^{k-1}, & C_k^2\mu_i^{k-2} & \vdots \\ & & \mu_i^k, & k\mu_i^{k-1} & C_k^2\mu_i^{k-2} \\ & & & \ddots & k\mu_i^{k-1} \\ \boldsymbol{0} & & & & \mu_i^k \end{bmatrix} \tag{8.2.11}$$

其中

$$C_k^j=\frac{k!}{j!\,(k-j)!}$$

如果 $S(\boldsymbol{G})<1$，那么所有 μ_i 的模均应小于 1。此时容易得知，对于任意与 k 无关的整数 s 均有：

$$\lim_{k\to\infty} C_k^s\cdot\mu_i^{k-s}=0 \tag{8.2.12}$$

因而，从(8.2.11)，(8.2.12)我们有：

$$\lim_{k\to\infty}\boldsymbol{J}_i^k=0$$

这样一来，再从(8.2.10)就可推知：

$$\lim_{k\to\infty}\boldsymbol{G}^k=\boldsymbol{P}\cdot\begin{bmatrix} \lim\limits_{k\to\infty}\boldsymbol{J}_1^k & & \boldsymbol{0} \\ & \lim\limits_{k\to\infty}\boldsymbol{J}_2^k & \\ \boldsymbol{0} & & \lim\limits_{k\to\infty}\boldsymbol{J}_s^k \end{bmatrix}\cdot\boldsymbol{P}^{-1}=0$$

于是，定理 2.3 的充分性即得证明。

8.2.2 一阶线性定常迭代法

现在讨论实践中经常用到的一阶线性定常迭代法。我们先叙述其收敛性和收敛速度概念，然后较详细地讨论逐次松弛法的有关问题。

(一)收敛性和收敛速度

解线性代数方程组的一阶线性定常迭代法的一般形式可以写为：

$$\boldsymbol{u}^{(k+1)}=\boldsymbol{G}\cdot\boldsymbol{u}^{(k)}+\boldsymbol{d}\quad(k=0,\ 1,\ 2,\ \cdots)\tag{8.2.7}$$

如果由其定义的向量序列$\{\boldsymbol{u}^{(k)}\}$有极限，显然，其极限应该是下列方程式的解：

$$\boldsymbol{u}=\boldsymbol{G}\boldsymbol{u}+\boldsymbol{d}$$

或

$$(\boldsymbol{I}-\boldsymbol{G})\boldsymbol{u}=\boldsymbol{d}\tag{8.2.13}$$

当然，我们也要求其极限是方程组(8.2.1)的解，这样对于求解它才是有意义的。今后我们只讨论使得方程组(8.2.13)与(8.2.1)有相同解的迭代法(8.2.7)，并称这样的迭代法为"相容"于方程组(8.2.1)的迭代法。其中矩阵$\boldsymbol{G}$就是与该迭代法相应的迭代矩阵。

构造相容于方程组(8.2.1)的一阶线性定常迭代法，可按如下方式进行。

首先将(8.2.1)的系数矩阵$\boldsymbol{A}$分解为一个非奇矩阵$\boldsymbol{Q}$与另一矩阵$\boldsymbol{R}$之差：

$$\boldsymbol{A}=\boldsymbol{Q}-\boldsymbol{R}\tag{8.2.14}$$

于是(8.2.1)变为：

$$(\boldsymbol{Q}-\boldsymbol{R})\boldsymbol{u}=\boldsymbol{b}$$

这一方程组显然与如下方程组等价(即有相同的解答)：

$$\boldsymbol{u}=\boldsymbol{Q}^{-1}\cdot\boldsymbol{R}\boldsymbol{u}+\boldsymbol{Q}^{-1}\cdot\boldsymbol{b}\tag{8.2.15}$$

只需令$\boldsymbol{G}=\boldsymbol{Q}^{-1}\cdot\boldsymbol{R}$, $\boldsymbol{d}=\boldsymbol{Q}^{-1}\cdot\boldsymbol{b}$, 我们就得到一个相容于方程组(8.2.1)的一阶线性定常迭代法。

现在，我们引进一阶线性定常迭代法的收敛性和收敛速度概念。

定义 2.4　如果对于任取的初始向量$\boldsymbol{u}^{(0)}$，由一阶线性定常迭代法(8.2.7)所产生的向量序列$\{\boldsymbol{u}^{(k)}\}$都有相同的极限，并且其极限就是方程组(8.2.1)的唯一解$\boldsymbol{u}^{*}$，则称迭代法(8.2.7)为收敛的，否则称为不收敛的。

在什么条件下迭代法(8.2.7)才是收敛的？为了说明这一点，我们引入下列误差向量：

$$\varepsilon^{(k)}=\boldsymbol{u}^{(k)}-\boldsymbol{u}^{*}\tag{8.2.16}$$

其中$\boldsymbol{u}^{*}$为方程组(8.2.1)的准确解。

因为迭代法(8.2.7)是相容于方程组(8.2.1)的，(8.2.1)的解$\boldsymbol{u}^{*}$就应满足(8.2.13)：

$$\boldsymbol{u}^{*}=\boldsymbol{G}\boldsymbol{u}^{*}+\boldsymbol{d}$$

将上式与(8.2.7)相减即得：

$$\varepsilon^{(k+1)}=\boldsymbol{G}\cdot\varepsilon^{(k)}\tag{8.2.17}$$

即是说，每迭代一次后，准确解与近似解之间的误差向量就被左乘以迭代矩阵$\boldsymbol{G}$。所以，自然有：

$$\varepsilon^{(k)}=\boldsymbol{G}\cdot\varepsilon^{(k-1)}=\boldsymbol{G}^{2}\cdot\varepsilon^{(k-2)}=\cdots=\boldsymbol{G}^{k}\cdot\varepsilon^{(0)}\tag{8.2.18}$$

如果迭代法(8.2.7)是收敛的，根据定义 2.4，对于任意初始误差向量$\varepsilon^{(0)}$我们均有：

$$\lim_{k\to\infty}\varepsilon^{(k)}=\lim_{k\to\infty}\boldsymbol{G}^{k}\cdot\varepsilon^{(0)}=0$$

这只有在$\lim\limits_{k\to\infty}\boldsymbol{G}^{k}=0$时才成立。所以，由(8.2.7)收敛之假定即可推知$\lim\limits_{k\to\infty}\boldsymbol{G}^{k}=0$。反之，若$\lim\limits_{k\to\infty}\boldsymbol{G}^{k}=0$, 从(8.2.18)知对于任意初始值均有$\lim\limits_{k\to\infty}\varepsilon^{(k)}=0$, 即$\lim\limits_{k\to\infty}\boldsymbol{u}^{(k)}=\boldsymbol{u}^{*}$。于是，我们看到迭代法(8.2.7)收敛和$\lim\limits_{k\to\infty}\boldsymbol{G}^{k}=0$是等价的。再利用定理 2.3，我们就得到如下结果：

定理 2.4　一阶线性定常迭代法(8.2.7)收敛的充要条件是其迭代矩阵$\boldsymbol{G}$的谱半径小

于1。亦即:

$$S(\boldsymbol{G})<1 \tag{8.2.19}$$

下面再来讨论收敛速度问题。从一阶线性定常迭代法收敛性问题的讨论，我们知道近似解 $\boldsymbol{u}^{(k)}$ 与真解 $\boldsymbol{u}^*$ 间的误差向量 $\varepsilon^{(k)}$ 是按(8.2.18)变化的。对此式两边取范数，就可得到:

$$\|\varepsilon^{(k)}\| \leqslant \|\boldsymbol{G}^k\| \cdot \|\varepsilon^{(0)}\|$$

如果迭代法(8.2.7)是收敛的,于是 $\lim\limits_{k\to\infty}\boldsymbol{G}^k=0$。根据定理2.2,我们总可以找到一个充分大的 k，使得 $\|\boldsymbol{G}^k\|$ 任意地小。这时，$\|\boldsymbol{G}^k\|$ 的大小将决定向量 $\varepsilon^{(k)}$ 收敛于零向量的速度。自然，我们应取与 $\|\boldsymbol{G}^k\|$ 有关的某个量作为收敛速度的一种度量。如果要求 $\|\varepsilon^{(k)}\|$ 减小为 $\|\varepsilon^{(0)}\|$ 的 ρ 倍($\rho<1$),即要求:

$$\|\varepsilon^{(k)}\| \leqslant \rho\|\varepsilon^{(0)}\|$$

为此仅需要

$$\|\boldsymbol{G}^{(k)}\| \leqslant \rho \quad \text{或者} \quad ((\|\boldsymbol{G}^k\|)^{\frac{1}{k}})^k \leqslant \rho$$

对此式两边取对数即可知迭代次数 k 应满足如下不等式:

$$k > -\lg\rho \Big/ \left(-\frac{1}{k}\lg\|\boldsymbol{G}^k\|\right) \tag{8.2.20}$$

可见为达到某一精度要求(由因子 ρ 表示),所需要的最少迭代次数是与量 $-\dfrac{1}{k}\lg\|\boldsymbol{G}^k\|$ 成反比的。自然,我们就把这个量定义为迭代过程(8.2.7)的平均收敛速度 $R_k(\boldsymbol{G})$:

$$R_k(\boldsymbol{G}) = -\frac{1}{k}\lg\|\boldsymbol{G}^k\| \tag{8.2.21}$$

上面所定义的平均收敛速度是难以具体算出来的，因而使用起来很不方便。实际计算中,常常用下面的办法来粗略地估计收敛速度。

假定迭代矩阵 $\boldsymbol{G}$ 的特征向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 是线性无关的，并以 $\mu_1, \mu_2, \cdots, \mu_n$ 表示其相应特征值,我们便可将初始误差向量 $\varepsilon^{(0)}$ 在这些特征向量上展开:

$$\varepsilon^{(0)} = \sum_{i=1}^{n} c_i \boldsymbol{x}_i$$

那么,从(8.2.18)容易得知:

$$\varepsilon^{(M)} = \boldsymbol{G}^M \cdot \varepsilon^{(0)} = \sum_{i=1}^{n} \mu_i^M c_i \boldsymbol{x}_i$$

如果按模最大的 μ_i 为单根,显然,当 M 很大时,上式中仅有与 $\max\limits_{1<i<n}|\mu_i| = S(\boldsymbol{G})$ 相应的一项占主导地位,我们假定就是其中的第一项。这样便有

$$\varepsilon^{(M)} \sim \mu_1^M \cdot c_1 \cdot \boldsymbol{x}_1$$

或者有

$$\|\varepsilon^{(k+M)}\| \sim S(\boldsymbol{G}^k)\|\varepsilon^{(M)}\|$$

即是说,当 M 很大时,如果我们要求误差向量 $\varepsilon^{(M)}$ 的模减小 ρ 倍,就只需要

$$S(\boldsymbol{G}^k) \leqslant \rho \quad \text{或者} \quad k \geqslant -\lg\rho / -\lg S(\boldsymbol{G}) \tag{8.2.22}$$

成立即可。(8.2.22)中后一不等式说明需要的迭代次数 k 与量

$$R(\boldsymbol{G}) = -\lg S(\boldsymbol{G}) \tag{8.2.23}$$

成反比例。因而，我们亦可用 $R(\boldsymbol{G})$ 作为收敛速度的一种度量。由于其仅在 k 充分大时代

表收敛速度的大小，同时，也由于可以证明(参见[19] p. 87)：

$$\lim_{k\to\infty}(\|\boldsymbol{G}^k\|)^{1/k}=S(\boldsymbol{G})$$

所以，通常将 $R(\boldsymbol{G})$ 叫作渐近收敛速度。有时为了简单起见，亦称其为收敛速度。

估计所需的迭代次数的最可靠的公式是(8.2.20)，但是由于其难于具体计算，我们还是经常用(8.2.22)来粗略估计，不过，有时会与实际情况相差较多。特别是矩阵 $\boldsymbol{G}$ 按模最大特征值相应的若当块之阶数大于1时，所需迭代次数比这样估计的次数往往要大得多，这一点应予注意。

(二)逐次松弛法及其收敛性

(1) 逐次松弛法的计算公式

现在讨论一种最常用的一阶线性定常迭代法——逐次松弛法。为了说明方便，仍以三阶方程组(8.2.2)为例。

最简单的一阶线性定常迭代法是简单迭代法，有时也叫作雅可比(Jacobi)迭代法。8.2.1节中已经得到了简单迭代法的计算公式(8.2.6)。(8.2.6)式也很容易从前面构造相容迭代格式的公式(8.2.14)和(8.2.15)直接导出。为此，只需将矩阵 $\boldsymbol{A}$ 分解为：

$$\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{C}$$

其中 $\boldsymbol{C}$ 为矩阵 $\boldsymbol{A}$ 的非对角线元素所构成的矩阵反号。

令(8.2.14)中的 $\boldsymbol{Q}=\boldsymbol{D}$, $\boldsymbol{R}=\boldsymbol{C}$, 然后用(8.2.15)式，我们立即得出简单迭代法的计算格式(8.2.6)。

简单迭代法(8.2.5)中，计算新的近似解分量 $u_i^{(k+1)}$ 时要用到前一次近似解 $\boldsymbol{u}^{(k)}$ 的所有分量(除 $u_i^{(k)}$ 外)，所以，需要同时存储 $\boldsymbol{u}^{(k+1)}$ 和 $\boldsymbol{u}^{(k)}$ 计算才能进行下去。为了节省存储量，可用(8.2.5)中第一式求得的近似解分量 $u_1^{(k+1)}$ 来代替 $u_1^{(k)}$ 代入第二个方程右端去求 $u_2^{(k+1)}$。然后将 $u_1^{(k+1)}$、$u_2^{(k+1)}$ 代入第三个方程右端去求 $u_3^{(k+1)}$，等等。这样，其迭代公式将为：

$$\begin{cases} u_1^{(k+1)}= & b_{12}u_2^{(k)}+b_{13}u_3^{(k)}+c_1 \\ u_2^{(k+1)}=b_{21}u_1^{(k+1)} & +b_{23}u_3^{(k)}+c_2 \\ u_3^{(k+1)}=b_{31}u_1^{(k+1)}+b_{32}u_2^{(k+1)} & +c_3 \end{cases}$$

这就是通常所谓高斯-赛德尔(Gauss-Seidel)迭代法的计算公式。

如果以 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 分别表示由矩阵 $\boldsymbol{C}$ 的下三角和上三角部分元素所构成的严格下和上三角型矩阵，即 $\boldsymbol{C}=\boldsymbol{L}+\boldsymbol{U}$, 那么，高斯-赛德尔迭代法的计算公式就可写为如下矩阵形式：

$$(\boldsymbol{D}-\boldsymbol{L})\boldsymbol{u}^{(k+1)}=\boldsymbol{U}\boldsymbol{u}^{(k)}+\boldsymbol{b}$$

或者

$$\boldsymbol{u}^{(k+1)}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\cdot\boldsymbol{U}\boldsymbol{u}^{(k)}+(\boldsymbol{D}-\boldsymbol{L})^{-1}\cdot\boldsymbol{b} \tag{8.2.24}$$

同样，若将矩阵 $\boldsymbol{A}$ 分解为：

$$\boldsymbol{A}=\boldsymbol{Q}-\boldsymbol{R}=(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}$$

按照前述相容迭代格式的构成原则(8.2.14)和(8.2.15)式，我们立即可以得到高斯-赛德尔迭代法的计算公式(8.2.24)。

逐次松弛法的计算公式与高斯-赛德尔迭代法很相似，不同之处仅在于引进了一个加速迭代收敛的参数，即所谓松弛因子 ω，而将迭代公式变为如下形式：

$$\begin{cases} u_1^{(k+1)}=\omega(\qquad\qquad b_{12}u_2^{(k)}+b_{13}u_3^{(k)}+c_1)+(1-\omega)u_1^{(k)} \\ u_2^{(k+1)}=\omega(b_{21}u_1^{(k+1)} \qquad\qquad +b_{23}u_3^{(k)}+c_2)+(1-\omega)u_2^{(k)} \\ u_3^{(k+1)}=\omega(b_{31}u_1^{(k+1)}+b_{32}u_2^{(k+1)} \qquad\qquad +c_3)+(1-\omega)u_3^{(k)} \end{cases} \tag{8.2.25}$$

或者

$$\begin{cases} u_1^{(k+1)}=\dfrac{\omega}{a_{11}}(b_1-a_{12}u_2^{(k)}-a_{13}u_3^{(k)})+(1-\omega)u_1^{(k)} \\ u_2^{(k+1)}=\dfrac{\omega}{a_{22}}(b_2-a_{21}u_1^{(k+1)}-a_{23}u_3^{(k)})+(1-\omega)u_2^{(k)} \\ u_3^{(k+1)}=\dfrac{\omega}{a_{33}}(b_3-a_{31}u_1^{(k+1)}-a_{32}u_2^{(k+1)})+(1-\omega)u_3^{(k)} \end{cases}$$

显然，当 $\omega=1$ 时，逐次松弛法的计算公式(8.2.25)就变为高斯-赛德尔迭代法。因而，适当地选取松弛因子 ω，逐次松弛法总可以得到不低于高斯-赛德尔迭代法的收敛速度。当$\omega>1$时，通常称为逐次超松弛法，$\omega<1$ 时，称为逐次低松弛法。容易看出，对于一般的 n 阶方程组，逐次松弛法计算公式应为：

$$u_i^{(k+1)}=\frac{\omega}{a_{ii}}\left\{b_i-\sum_{j=1}^{i-1}a_{ij}u_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}u_j^{(k)}\right\}+(1-\omega)u_i^{(k)} \tag{8.2.26}$$

$$(i=1,\ 2,\ \cdots,\ n)$$

如果仍旧采用前面的矩阵符号并将(8.2.26)中 $\boldsymbol{u}^{(k+1)}$ 解出，即可把逐次松弛法的迭代公式表为如下矩阵形式：

$$\boldsymbol{u}^{(k+1)}=\omega\cdot(\boldsymbol{D}^{-1}\cdot\boldsymbol{L}\boldsymbol{u}^{(k+1)}+\boldsymbol{D}^{-1}\cdot\boldsymbol{U}\boldsymbol{u}^{(k)}+\boldsymbol{D}^{-1}\cdot\boldsymbol{b})+(1-\omega)\boldsymbol{u}^{(k)}$$

或者：

$$\begin{aligned} \boldsymbol{u}^{(k+1)}&=\boldsymbol{L}_\omega\cdot\boldsymbol{u}^{(k)}+(\boldsymbol{D}-\omega\boldsymbol{L})^{-1}\cdot\omega\cdot\boldsymbol{b} \\ \boldsymbol{L}_\omega&=(\boldsymbol{D}-\omega\boldsymbol{L})^{-1}\cdot(\omega\boldsymbol{U}+(1-\omega)\boldsymbol{D}) \end{aligned} \tag{8.2.27}$$

读者易于自己验证，若将矩阵 $\boldsymbol{A}$ 分解为如下形式：

$$\boldsymbol{A}=\boldsymbol{Q}-\boldsymbol{R}=\frac{1}{\omega}(\boldsymbol{D}-\omega\boldsymbol{L})-\frac{1}{\omega}((1-\omega)\boldsymbol{D}+\omega\boldsymbol{U})$$

按照前述建立相容迭代法的公式(8.2.14)和(8.2.15)，立即可以得到逐次松弛法的计算公式(8.2.27)。

简单迭代法由于收敛太慢，实践中很少使用。高斯-赛德尔迭代法则可视为逐次松弛法的特殊情况，后者在实践中是经常使用的。因而，下面我们主要讨论逐次松弛法。

(2) 逐次松弛法的收敛性问题。读者首先感兴趣的自然是松弛因子 ω 究竟应选取在什么范围内，逐次松弛法才有可能收敛，如下定理回答了这一问题：

定理 2.5 逐次松弛法(8.2.27)收敛的必要条件是：

$$0<\omega<2 \tag{8.2.28}$$

这一定理的证明比较简单。从定理 2.4 知，若要逐次松弛法(8.2.27)收敛，必须迭代矩阵 $\boldsymbol{L}_\omega$ 的谱半径 $S(\boldsymbol{L}_\omega)$ 小于1。然而，注意到矩阵的行列式值等于其特征值的乘积以及 $\boldsymbol{L}$，$\boldsymbol{U}$ 分别为严格的下和上三角矩阵，从(8.2.27)式我们有：

$$\begin{aligned} S(\boldsymbol{L}_\omega)\geqslant|\det(\boldsymbol{L}_\omega)|^{1/n}&=[|\det(\boldsymbol{D}-\omega\boldsymbol{L})^{-1}|\cdot|\det((1-\omega)\boldsymbol{D}+\omega\boldsymbol{U})|]^{1/n} \\ &=[|1-\omega|^n]^{1/n}=|1-\omega| \end{aligned}$$

故若要 $S(\boldsymbol{L}_\omega)<1$，则需 $0<\omega<2$。

上述定理说明，对于任何系数矩阵 $\boldsymbol{A}$，若要逐次松弛法(8.2.27)收敛，必须选取松弛因子 ω 为(0, 2)间的正数。然而，当松弛因子 ω 满足条件 $0<\omega<2$ 时，并不是对于所有的系数矩阵 $\boldsymbol{A}$ 来说，逐次松弛法(8.2.27)均是收敛的。目前已对许多类系数矩阵研究过逐次松弛法的收敛性问题，我们将简单地讨论一下其中某些有用结果。

由于对称正定矩阵在实践中是最常遇到的，例如，用有限元素法求解弹性结构的静力平衡问题，就归结为解一个对称正定(或半正定)系数矩阵的线性代数方程组。因而，我们首先来讨论矩阵 $\boldsymbol{A}$ 为对称正定的情形。

定理 2.6　若矩阵 $\boldsymbol{A}$ 为对称矩阵(或爱尔米特矩阵)且其对角线元素均为正实数，则逐次松弛法(8.2.27)当 $0<\omega<2$ 时收敛的充要条件是 $\boldsymbol{A}$ 为正定矩阵。

这一定理说明，对于对称正定矩阵，只要 $0<\omega<2$，逐次松弛法(8.2.27)总是收敛的，所以，对于实际使用松弛法来说，这个定理比较重要。下面，我们就来证明这个定理。

假定矩阵 $\boldsymbol{A}$ 是爱尔米特矩阵，即其共轭转置阵 $\boldsymbol{A}^*=\boldsymbol{A}$。由于其对角线元为正实数，故 $\boldsymbol{A}$ 有如下分解式：

$$\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{L}^* \tag{8.2.29}$$

其中 $\boldsymbol{L}$ 为矩阵 $\boldsymbol{A}$ 的下三角部分反号，$\boldsymbol{D}$ 为正定对角线型矩阵。

我们仍以 $\boldsymbol{\varepsilon}^{(k)}$ 表示按(8.2.27)迭代 k 次所得近似解 $\boldsymbol{u}^{(k)}$ 与真解 $\boldsymbol{u}^*$ 间的误差向量。从(8.2.27)式得知：

$$\boldsymbol{\varepsilon}^{(k+1)}=\boldsymbol{L}_\omega\cdot\boldsymbol{\varepsilon}^{(k)}\quad(k\geqslant0)$$

或者说(8.2.27)等价地有：

$$(\boldsymbol{D}-\omega\boldsymbol{L})\boldsymbol{\varepsilon}^{(k+1)}=(\omega\boldsymbol{L}^*+(1-\omega)\boldsymbol{D})\cdot\boldsymbol{\varepsilon}^{(k)}\quad(k\geqslant0)$$

令 $\boldsymbol{\delta}^{(k)}=\boldsymbol{\varepsilon}^{(k)}-\boldsymbol{\varepsilon}^{(k+1)}$，$(k\geqslant0)$，并以矩阵 $\boldsymbol{A}$ 的分解式(8.2.29)分别代入上式的右端和左端，我们就可以得到如下两个等式：

$$\left.\begin{array}{l}(\boldsymbol{D}-\omega\boldsymbol{L})\boldsymbol{\delta}^{(k)}=\omega\boldsymbol{A}\boldsymbol{\varepsilon}^{(k)}\\ \omega\boldsymbol{A}\boldsymbol{\varepsilon}^{(k+1)}=(1-\omega)\boldsymbol{D}\cdot\boldsymbol{\delta}^{(k)}+\omega\boldsymbol{L}^*\cdot\boldsymbol{\delta}^{(k)}\end{array}\right\}\quad(k\geqslant0) \tag{8.2.30}$$

再用 $\boldsymbol{\varepsilon}^{(k)}$ 及 $\boldsymbol{\varepsilon}^{(k+1)}$ 分别与(8.2.30)中两式的等式两端作内积，然后相减之即得：

$$\begin{aligned}&\omega\{(\boldsymbol{\varepsilon}^{(k)},\ \boldsymbol{A}\boldsymbol{\varepsilon}^{(k)})-(\boldsymbol{\varepsilon}^{(k+1)},\ \boldsymbol{A}\boldsymbol{\varepsilon}^{(k+1)})\}\\&=(\boldsymbol{\varepsilon}^{(k)},\ (\boldsymbol{D}-\omega\boldsymbol{L})\boldsymbol{\delta}^{(k)})-(1-\omega)(\boldsymbol{\varepsilon}^{(k+1)},\ \boldsymbol{D}\boldsymbol{\delta}^{(k)})-\omega(\boldsymbol{\varepsilon}^{(k+1)},\ \boldsymbol{L}^*\boldsymbol{\delta}^{(k)})\\&=(\boldsymbol{\delta}^{(k)},\ \boldsymbol{D}\boldsymbol{\delta}^{(k)})-\omega(\boldsymbol{\delta}^{(k)},\ \boldsymbol{L}\boldsymbol{\delta}^{(k)})+(\omega\boldsymbol{A}\boldsymbol{\varepsilon}^{(k+1)},\ \boldsymbol{\delta}^{(k)})\end{aligned}$$

再将(8.2.30)的第2式代入上式，即得：

$$\begin{aligned}&\omega\{(\boldsymbol{\varepsilon}^{(k)},\ \boldsymbol{A}\boldsymbol{\varepsilon}^{(k)})-(\boldsymbol{\varepsilon}^{(k+1)},\ \boldsymbol{A}\boldsymbol{\varepsilon}^{(k+1)})\}\\&=(\boldsymbol{\delta}^{(k)},\ \boldsymbol{D}\boldsymbol{\delta}^{(k)})-\omega(\boldsymbol{\delta}^{(k)},\ \boldsymbol{L}\boldsymbol{\delta}^{(k)})+(1-\omega)(\boldsymbol{\delta}^{(k)},\ \boldsymbol{D}\boldsymbol{\delta}^{(k)})+\omega(\boldsymbol{\delta}^{(k)},\ \boldsymbol{L}\boldsymbol{\delta}^{(k)})\\&=(2-\omega)\cdot(\boldsymbol{\delta}^{(k)},\ \boldsymbol{D}\boldsymbol{\delta}^{(k)})\quad(k\geqslant0)\end{aligned} \tag{8.2.31}$$

利用关系式(8.2.31)，我们就可证明定理2.6。首先，假定 $\boldsymbol{A}$ 是正定的，且 $0<\omega<2$。我们可以选取 $\boldsymbol{\varepsilon}^{(0)}$ 为迭代矩阵 $\boldsymbol{L}_\omega$ 的任一特征向量，将其相应特征值记为 λ，于是有 $\boldsymbol{\varepsilon}^{(1)}=\lambda\boldsymbol{\varepsilon}^{(0)}$，$\boldsymbol{\delta}^{(0)}=(1-\lambda)\boldsymbol{\varepsilon}^{(0)}$。同时，将其代入关系式(8.2.31)便有：

$$\left(\frac{2-\omega}{\omega}\right)|1-\lambda|^2(\boldsymbol{\varepsilon}^{(0)},\ \boldsymbol{D}\boldsymbol{\varepsilon}^{(0)})=(1-|\lambda|^2)\cdot(\boldsymbol{\varepsilon}^{(0)},\ \boldsymbol{A}\cdot\boldsymbol{\varepsilon}^{(0)}) \tag{8.2.32}$$

于是从 $\boldsymbol{A}$ 的正定性有 $\boldsymbol{A}\cdot\boldsymbol{\varepsilon}^{(0)}\neq0$，再从(8.2.30)式即得知 $\boldsymbol{\delta}^{(0)}=\omega(\boldsymbol{D}-\omega\cdot\boldsymbol{L})^{-1}\cdot\boldsymbol{A}\boldsymbol{\varepsilon}^{(0)}\neq0$。这样 $\lambda\neq1$。故当 $0<\omega<2$ 时，(8.2.32)两端均应为正数。这只有 $|\lambda|<1$ 时才是可能的，即

是说逐次松弛法必收敛。

其次，若假定逐次松弛法收敛，则从定理 2.5 应有：$0<\omega<2$。同时，对于任意初始向量 $\varepsilon^{(0)}$ 均有 $\varepsilon^{(k)}\to 0(k\to\infty)$。又因 $\boldsymbol{D}$ 是正定矩阵，所以，从(8.2.31)式可知：

$$(\varepsilon^{(k)},\ \boldsymbol{A}\varepsilon^{(k)})=(\varepsilon^{(k+1)},\ \boldsymbol{A}\varepsilon^{(k+1)})+\left(\frac{2-\omega}{\omega}\right)\cdot(\delta^{(k)},\ \boldsymbol{D}\delta^{(k)})$$
$$\geqslant(\varepsilon^{(k+1)},\ \boldsymbol{A}\cdot\varepsilon^{(k+1)})\quad(k\geqslant 0)\tag{8.2.33}$$

如果矩阵 $\boldsymbol{A}$ 不是正定的，我们必可找到向量 $\varepsilon^{(0)}$，使得 $(\varepsilon^{(0)},\ \boldsymbol{A}\varepsilon^{(0)})\leqslant 0$。同时，由于 $\boldsymbol{L}_\omega$ 的特征值之模均小于 1，$\delta^{(0)}=(1-\boldsymbol{L}_\omega)\varepsilon^{(0)}$ 应该为非零向量，这样，我们便有：

$$(\varepsilon^{(1)},\ \boldsymbol{A}\varepsilon^{(1)})<(\varepsilon^{(0)},\ \boldsymbol{A}\varepsilon^{(0)})\leqslant 0$$

再利用(8.2.33)式中的不增性质，我们必定有：

$$(\varepsilon^{(k)},\ \boldsymbol{A}\varepsilon^{(k)})<(\varepsilon^{(0)},\ \boldsymbol{A}\varepsilon^{(0)})\leqslant 0$$

这一事实显然与 $\varepsilon^{(k)}\to 0$ 相矛盾，故矩阵 $\boldsymbol{A}$ 应为正定矩阵，于是定理 2.6 得证。

除对称正定矩阵外，已对许多其它类型的系数矩阵研究过逐次松弛法的收敛性问题。我们将其中比较常见的两种情况叙述如下，以供读者参考。

定理 2.7　若矩阵 $\boldsymbol{A}$ 满足如下两个条件，则当 $0<\omega\leqslant 1$ 时逐次松弛法(8.2.27)必收敛。

条件(1)　找不到排列矩阵 $\boldsymbol{P}$，使得：

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}=\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ 0 & \boldsymbol{A}_{22}\end{pmatrix}\tag{8.2.34}$$

其中 $\boldsymbol{A}_{11}$，$\boldsymbol{A}_{22}$ 为方阵。

条件(2)
$$|a_{ii}|\geqslant\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\quad(i=1,\ 2,\ \cdots,\ n)\tag{8.2.35}$$

且至少对于某一个 i 值，上式中出现严格的不等号。

满足第一个条件的矩阵称为不可约(分解)矩阵，其意义为相应的方程组不能化为两个相互独立的方程组来分别求解。满足第二个条件的矩阵称为弱对角优势矩阵，因而定理也可叙述为：对于不可约的弱对角优势矩阵，逐次松弛法(8.2.27)当 $0<\omega\leqslant 1$ 时必定是收敛的。

定理 2.8　假如分解式 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$ 中，矩阵 $\boldsymbol{D}$、$\boldsymbol{L}$、$\boldsymbol{U}$ 均为非负矩阵(即其每个元素均大于等于零)，其中 $\boldsymbol{D}$ 为非奇异对角型矩阵，$\boldsymbol{L}$ 及 $\boldsymbol{U}$ 分别为严格下和上三角型矩阵，则逐次松弛法(8.2.27)收敛的充分条件为矩阵 $\boldsymbol{D}^{-1}\cdot(\boldsymbol{L}+\boldsymbol{U})$ 之谱半径 $S(\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U}))$ 小于 1 且 $0<\omega\leqslant 1$。

关于定理 2.7、2.8 的证明，可参阅[13]和[19]，这里不再进行讨论。

尽管目前对于逐次松弛法的收敛性问题已进行了许多研究，但是应该说逐次松弛法对于任意系数矩阵情况的收敛性问题，尚未得到彻底解决。不过实际应用中绝大多数系数矩阵均是对称正定矩阵，根据定理 2.6，其收敛性总是有保证的。对于这些问题重要的是如何来选择加速迭代过程收敛的松弛因子 ω，以节省计算工作量的问题。我们将在下一段中讨论这个问题。

(三)松弛因子的选取问题

本节讨论所谓最优松弛因子的选取问题。使得逐次松弛法的渐近收敛速度最快的松弛因子通常称为最优松弛因子，并用 ω_{opt} 记之。对于一般矩阵(即使是对称正定矩阵)，目前尚

无确定 ω_{opt} 的理论结果。实际计算时，大都由计算经验或通过试算来定出 ω_{opt} 的近似值。仅对某些特殊类型的矩阵，例如所谓有性质"A"的矩阵，有确定 ω_{opt} 的理论公式。下面，我们先对有性质"A"的矩阵来讨论最优松弛因子的选取问题，然后不加证明地介绍其它一些理论结果，以便读者参考。最后讨论一下最优松弛因子的试算确定问题。

(1) 性质"A"和相容次序。讨论松弛因子选取问题时，要用到性质"A"和相容次序的概念，我们先来简单叙述一下这些概念。

定义 2.5 若矩阵 $\boldsymbol{A}$ 具如下形式：

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{U}_1 & & & \\ \boldsymbol{L}_2 & \boldsymbol{D}_2 & \boldsymbol{U}_2 & & \boldsymbol{0} \\ & \boldsymbol{L}_3 & \boldsymbol{D}_3 & \boldsymbol{U}_3 & \\ & & \ddots & \ddots & \ddots \boldsymbol{U}_{m-1} \\ \boldsymbol{0} & & & \ddots \boldsymbol{L}_m & \ddots \boldsymbol{D}_m \end{bmatrix} \tag{8.2.36}$$

其中 $\boldsymbol{D}_i$ 为对角线型方阵，$\boldsymbol{L}_i$、$\boldsymbol{U}_i$ 为相应阶数的长方阵，则称矩阵 $\boldsymbol{A}$ 为 D-型分块三对角线矩阵。

形如(8.2.36)的 D-型分块三对角线矩阵，是下列方式定义的具有"性质 A"的矩阵(亦称二循环矩阵)的特殊情况。

定义 2.6 若能将前 n 个正整数所构成的集合 $W=\{1, 2, \cdots, n\}$ 分为两个不相交的子集合 S_1, S_2(即 $S_1+S_2=W$, S_1 与 S_2 无公共元素)，使得矩阵 $\boldsymbol{A}$ 对角线以外的每一个非零元素 $a_{ij}(i\neq j)$ 的足标对 (i, j)，均满足条件：$i\in S_1$, $j\in S_2$, 或者 $j\in S_1$, $i\in S_2$, 则称此矩阵具有"性质 A"。

实际上，我们将具有"性质 A"的矩阵 $\boldsymbol{A}$ 中编号属于集合 S_1 的各行(以及相应的各列)均排至前面，而将属于集合 S_2 的均排至后面，就可把矩阵 $\boldsymbol{A}$ 变为如下形式：

$$\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}=\begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{H} \\ \boldsymbol{K} & \boldsymbol{D}_2 \end{bmatrix}$$

其中，$\boldsymbol{D}_1$、$\boldsymbol{D}_2$ 为对角线型方阵；$\boldsymbol{P}$ 为相应的排列矩阵。显然，这是(8.2.36)的一种特殊情况。因而，具有"性质 A"的矩阵总可以通过行列的同时排列，变为(8.2.36)的形状。反之，若将(8.2.36)中与对角线子块 $\boldsymbol{D}_{2k}(k=1, 2, \cdots)$ 的行编号相应的 W 的子集记为 S_1, 与 $\boldsymbol{D}_{2k-1}$ 的行编号相应的 W 的子集记为 S_2, 则很容易按照定义 2.6 验证形如(8.2.36)的矩阵具有"性质 A"。

从上述讨论得知，形如(8.2.36)的矩阵与具有"性质 A"的矩阵仅仅是行列的次序不同。但是逐次松弛法的计算格式(8.2.26)是与方程式和未知数的排列次序有关系的。讨论最优松弛因子的选取时，为确定起见，应该首先规定方程式和未知数次序的排列办法，同时，应找出什么样的次序是比较好的。已经证明，所谓"相容次序"是比较合理的次序。"相容次序"的定义如下：

定义 2.7 若能将前 n 个正整数所构成的集合 W 分为 t 个不相交的子集 $S_1, S_2, \cdots, S_t$(即 $\sum_{k=1}^{t} S_k=W$, S_i 与 $S_j(i\neq j)$ 无公共元素)，使得矩阵 $\boldsymbol{A}$ 任意非零非对角线元 $a_{ij}\neq 0$ $(i\neq j)$ 的足标对 (i, j) 满足如下条件：若 $i\in S_k$ 则 $j\in S_{k-1}$(当 $j<i$) 或 $j\in S_{k+1}$(当 $j>i$)，则称此矩阵具有"相容次序"。

如果将(8.2.36)中属于对角线子块 D_k 之行编号记为集合 S_k，显然，$\sum_{k=1}^{m} S_k = W$ 且 S_i 与 $S_j(i\neq j)$ 无公共元素。很容易验证(8.2.36)中矩阵的任意非零非对角线元素 $a_{ij}\neq 0(i\neq j)$，其足标对(i, j)均满足定义 2.7 中的条件。所以，形如(8.2.36)的矩阵是具有“相容次序”的。

(2) 最优松弛因子的理论计算公式。

上面的讨论说明，形如(8.2.36)的矩阵是具有“性质 A”和“相容次序”矩阵的一种特殊情况。由于其形状较为简单，我们就以它为例来讨论最优松弛因子 ω_{opt} 的选取问题，但是结论对于其它具有“性质 A”与“相容次序”的矩阵也是成立的。同时，由于实践中最常遇到对称正定矩阵，故下面仅就对称正定且形如(8.2.36)的矩阵来讨论其最优松弛因子的理论计算公式问题。

定理 2.9　假定方程组(8.2.1)的系数矩阵 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$ 为对称正定矩阵且具有(8.2.36)的形状，其中 $\boldsymbol{D}$ 为对角矩阵，$\boldsymbol{L}$ 和 $\boldsymbol{U}=\boldsymbol{L}^T$ 分别为严格下与上三角形矩阵。则使逐次松弛法(8.2.27)的渐近收敛速度最快的最优松弛因子为：

$$\omega_{opt}=\frac{2}{1+\sqrt{1-\mu_1^2}} \tag{8.2.37}$$

其中 $\mu_1(|\mu_1|<1)$ 为矩阵 $\boldsymbol{D}^{-1}\cdot(\boldsymbol{L}+\boldsymbol{U})$ 的按模最大特征值。

证明：由于 $\boldsymbol{D}$ 的正定性，$\boldsymbol{D}^{-1/2}$ 存在。矩阵 $\boldsymbol{D}^{-1}\cdot(\boldsymbol{L}+\boldsymbol{U})$ 与矩阵 $\boldsymbol{B}=\boldsymbol{D}^{-1/2}\cdot(\boldsymbol{L}+\boldsymbol{U})\cdot\boldsymbol{D}^{-1/2}$ 是相似的，两者应有相同特征值 μ_i。再从矩阵 $\boldsymbol{A}$ 的对称性可知 $\boldsymbol{U}=\boldsymbol{L}^T$，故得知矩阵 $\boldsymbol{B}$，因而矩阵 $\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$ 的特征值 μ_i 均是实数。另外，矩阵 $\boldsymbol{A}$ 有形状(8.2.36)，所以 μ_i 应为如下多项式的零点：

$$P(\mu)=\det(-\boldsymbol{L}+\mu\boldsymbol{D}-\boldsymbol{U})=\begin{vmatrix} \mu\boldsymbol{D}_1 & \boldsymbol{U}_1 & & & \\ \boldsymbol{L}_2 & \mu\boldsymbol{D}_2 & \boldsymbol{U}_2 & & \boldsymbol{0} \\ & \boldsymbol{L}_3 & \mu_3\boldsymbol{D} & \boldsymbol{U}_3 & \\ & & \ddots & \ddots & \boldsymbol{U}_{m-1} \\ \boldsymbol{0} & & & \boldsymbol{L}_m & \mu\boldsymbol{D}_m \end{vmatrix} \tag{8.2.38}$$

如果将此行列式中处于子块 $\boldsymbol{D}_{2k-1}(k=1, 2, \cdots)$ 处的所有行和列同时乘以 -1，显然，行列式之值不变。这样，将得到：

$$P(\mu)=\begin{vmatrix} \mu\boldsymbol{D}_1 & -\boldsymbol{U}_1 & & & \\ -\boldsymbol{L}_2 & \mu\boldsymbol{D}_2 & -\boldsymbol{U}_2 & & \\ & -\boldsymbol{L}_3 & \mu\boldsymbol{D}_3 & -\boldsymbol{U}_3 & \\ & & \ddots & \ddots & -\boldsymbol{U}_{m-1} \\ & & & -\boldsymbol{L}_m & \mu\boldsymbol{D}_m \end{vmatrix}$$

上式中以 $-\mu$ 代替 μ，并提出公因子“-1”即得：

$$P(-\mu)=(-1)^n P(\mu)$$

由此可知，$P(\mu)$ 的非零零点是成对出现的，每一对零点之绝对值相等，符号相反。

另外，逐次松弛法(8.2.27)的迭代矩阵可写为：

$$\boldsymbol{L}_\omega=\left(\frac{1}{\omega}\boldsymbol{D}-\boldsymbol{L}\right)^{-1}\cdot\left(\boldsymbol{U}+\left(\frac{1}{\omega}-1\right)\boldsymbol{D}\right)$$

所以，其特征值 λ 应为如下多项式的零点：

$$Q(\lambda)=\det\left\{(\omega^{-1}\boldsymbol{D}-\boldsymbol{L})\lambda-\left(\boldsymbol{U}+\left(\frac{1}{\omega}-1\right)\boldsymbol{D}\right)\right\}$$
$$=\det(-\boldsymbol{L}\lambda+\omega^{-1}(\lambda+\omega-1)\boldsymbol{D}-\boldsymbol{U})$$

若令:
$$\omega^{-1}\cdot(\lambda+\omega-1)=\xi$$

并考虑到矩阵 $\boldsymbol{A}$ 的形状(8.2.36),我们有:

$$Q(\lambda)=\begin{vmatrix} \xi\boldsymbol{D}_1 & \boldsymbol{U}_1 & & & \\ \lambda\boldsymbol{L}_2 & \xi\boldsymbol{D}_2 & \boldsymbol{U}_2 & & \boldsymbol{0} \\ & \ddots & \ddots & \ddots & \\ & & \lambda\boldsymbol{L}_{m-1} & \xi\boldsymbol{D}_{m-1} & \boldsymbol{U}_{m-1} \\ \boldsymbol{0} & & & \lambda\boldsymbol{L}_m & \xi\boldsymbol{D}_m \end{vmatrix};$$

将上述行列式中处于子块 $\boldsymbol{D}_k(k=1,\ 2,\ \cdots,\ n)$ 所在位置的各列乘以因子 $\lambda^{\left(\frac{k}{2}-1\right)}$，各行乘以因子 $\lambda^{(1-k)/2}$,即可消去 $\boldsymbol{L}_i$ 前面的因子 λ,再考虑到多一个因子 $\lambda^{\frac{n}{2}}$,便得到:

$$Q(\lambda)=\lambda^{\frac{n}{2}}\cdot\begin{vmatrix} \xi\lambda^{-\frac{1}{2}}\boldsymbol{D}_1, & \boldsymbol{U}_1 & \boldsymbol{0} & \\ \boldsymbol{L}_2, & \xi\lambda^{-\frac{1}{2}}\boldsymbol{D}_2, & \boldsymbol{U}_2 & \\ & \ddots & \ddots & \boldsymbol{U}_{m-1} \\ \boldsymbol{0} & & \boldsymbol{L}_m, & \xi\lambda^{-\frac{1}{2}}\boldsymbol{D}_m \end{vmatrix}$$

将此式与(8.2.38)式对照之,我们得到:

$$Q(\lambda)=\lambda^{\frac{n}{2}}P(\xi\lambda^{-\frac{1}{2}})=\lambda^{\frac{n}{2}}P(\lambda^{-\frac{1}{2}}\cdot\omega^{-1}\cdot(\lambda+\omega-1))$$

由于 $0<\omega<2$ 且 $P(\mu)$ 的零点成对出现,如果 μ 是 $P(\mu)$ 的任意零点,那么,满足下列关系式的任意 λ 值亦应为 $Q(\lambda)$ 之零点:

$$\frac{(\lambda+\omega-1)^2}{\lambda}=\omega^2\mu^2 \tag{8.2.39}$$

反之也成立。

不难看出,上述关系式的推导中并未真正用到矩阵 $\boldsymbol{A}$ 的对称性,所以,这一关系式对于任意形状为(8.2.36)的矩阵也成立。当矩阵 $\boldsymbol{A}$ 是对称正定矩阵时,前面已证明逐次松弛法的收敛性,所以 $|\lambda|<1$。在上式中令 $\omega=1$,即可得知 $\mu^2=\lambda$,故 $|\mu|<1$。此外,μ_i 亦应为实数且成对出现等。此时,由关系式(8.2.39)所建立的 μ_i 与 λ_i 间的对应关系(当 $\omega>1$ 且给定),就可用图 8.8 定性地描述:

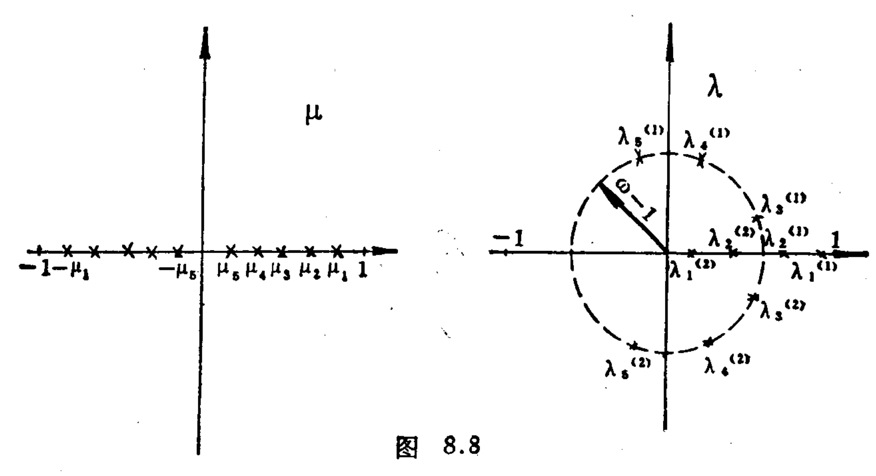


图 8.8

由于对给定的 μ_i, λ_i 是下列二次方程的根:

$$\lambda^2+[2(\omega-1)-\omega^2\mu_i^2]\lambda+(\omega-1)^2=0 \tag{8.2.40}$$

故 $\mu_i=0$ 对应于 $\lambda_i=-(\omega-1)$。每一对非零特征值 $\pm\mu_i\neq0$, 对应于两个值 $\lambda_i^{(1)}$、$\lambda_i^{(2)}$, 其乘积等于 $(\omega-1)^2$。若 $\lambda_i^{(1)}$ 为复数则必有 $\lambda_i^{(2)}=\bar{\lambda}_i^{(1)}$, 两者之模均为 $\omega-1$, 此模不依赖于相应的 μ_i 值。所以 λ_i 在 λ-平面上相对于半径为 $\omega-1$ 的圆来说, 应有如图 8.8 的配置。显然, 若 λ_i 中有实数者, 按模最大的 λ_i 必为实数, 且其模大于 $\omega-1$, 例如, 图中的 $\lambda_1^{(1)}$。并且, 其相应的 μ 值必为 $\pm\mu_1$。

详细地分析前述二次方程(8.2.40)(即将其两端对 ω 微商), 得知其按模最大根 $\lambda_1^{(1)}$ 对 ω 的导数为:

$$\frac{d\lambda_1^{(1)}}{d\omega}=-\frac{\lambda_1^{(1)}-\lambda_1^{(1)}\omega\mu_1^2+\omega-1}{\lambda_1^{(1)}-\dfrac{1}{2}\omega^2\mu_1^2+\omega-1}$$

当 $\omega=1$ 时, 因为 $\lambda_1^{(1)}=\mu_1^2<1$ 且为正实数, 所以,

$$\left.\frac{d\lambda_1^{(1)}}{d\omega}\right|_{\omega=1}=-2(1-\mu_1^2)<0$$

即是说, $\lambda_1^{(1)}$ 在 $\omega=1$ 处随 ω 之增加将减小其值。此外, $\dfrac{d\lambda_1^{(1)}}{d\omega}$ 本身亦应是 ω 的连续函数, 而其分子当 $1<\omega<2$ 时是不变号的(因其在 $\omega=(1-\lambda_1^{(1)})/(1-\lambda_1^{(1)}\mu_1^2)<1$ 时为零), 所以, 在区间 $1<\omega<2$ 内仅当其分母为零时 $d\lambda_1^{(1)}/d\omega$ 改变符号。这样, 当 ω 从 1 增加时, $\lambda_1^{(1)}$ 之值应减小, 但仍为正实数, 直到 $\lambda_1^{(1)}=\dfrac{1}{2}\omega^2\mu_1^2-\omega+1$ 为止, 此时, $\dfrac{d\lambda_1^{(1)}}{d\omega}$ 为无限大。直接代入二次方程(8.2.40)可知, 当 $\omega=\omega_s=\dfrac{2}{1+\sqrt{1-\mu_1^2}}$ 时, $\lambda_1^{(1)}=\dfrac{1}{2}\omega^2\mu_1^2-\omega+1$ 为其两重根, 其相应判别式为零, 此时所有的 $\lambda_i^{(1,2)}$ 均在半径等于 ω_s-1 的圆上。当 $\omega_s<\omega<2$ 时, $\lambda_1^{(1)}$ 将变为复根, 其模等于 $\omega-1>\omega_s-1$。当 $0<\omega<1$ 时, 类似的讨论可以得知 $S(\boldsymbol{L}_\omega)>S(\boldsymbol{L}_1)$。因而, 我们可以断定 ω_s 将使 $\lambda_1^{(1)}$ 之模最小, 即 $S(\boldsymbol{L}_\omega)$ 最小。这样, 最优松弛因子的计算公式应为:

$$\omega_{opt}=\omega_s=\frac{2}{1+\sqrt{1-\mu_1^2}}$$

此时, 逐次松弛法迭代矩阵的谱半径为:

$$S(\boldsymbol{L}_{\omega opt})=\frac{1}{2}\omega_{opt}^2\cdot\mu_1^2-\omega_{opt}+1=\omega_{opt}-1 \tag{8.2.41}$$

上述的松弛因子 ω 与 $S(\boldsymbol{L}_\omega)$ 间的关系, 可以定性地归纳于图 8.9 之中:

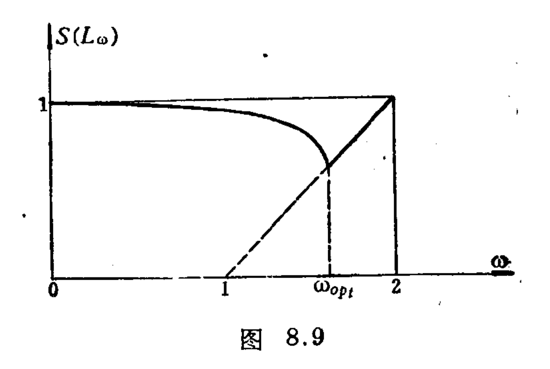


图 8.9

由图 8.9 可知，当 $\omega<\omega_{opt}$ 并逐渐增加趋近于 ω_{opt} 时，$S(\boldsymbol{L}_\omega)$ 曲线的切线趋近于铅垂线。但当 $\omega>\omega_{opt}$ 并逐渐减小趋近于 ω_{opt} 时，该切线方向不变，其斜率恒为 1。即是说 $\omega_{opt}-\Delta\omega$ 与 $\omega_{opt}+\Delta\omega$ 中，后者相应的谱半径较之 $\omega_{opt}-1$ 的增量更小一些。所以，取 ω 略大于 ω_{opt} 进行计算较取 ω 略小于 ω_{opt} 进行计算更为有利。这一点对于实际计算 ω_{opt} 的近似值时是重要的。

除上述具有性质"A"的矩阵有最优松弛因子的理论计算公式外，对于其它一些矩阵，还有一些计算最优松弛因子的近似公式。现将其中常用的两个公式不加证明地叙述如下，以供参考。

定理 2.10　若矩阵 $\boldsymbol{A}=\boldsymbol{I}-\boldsymbol{L}-\boldsymbol{L}^T$ 为对称正定矩阵（对角线元为 1，下三角部分为 $-\boldsymbol{L}$），其特征值为 $0<\lambda_n\leqslant\cdots\leqslant\cdots\leqslant\lambda_1$，矩阵 $\boldsymbol{H}=\boldsymbol{L}^T-\boldsymbol{L}$ 之谱半径为 $S(\boldsymbol{H})=\sigma$，则可按如下公式粗略地确定最优松弛因子：

$$\omega_{opt}\sim\omega^*=\begin{cases}\dfrac{2}{1+\sqrt{\lambda_1\cdot\lambda_n-\sigma^2}} & \text{当 } \lambda_n\cdot(\lambda_1-\lambda_n)>2\sigma^2\\[2ex] \dfrac{2}{1+\sqrt{\lambda_n^2+\sigma^2}} & \text{当 } \lambda_n\cdot(\lambda_1-\lambda_n)\leqslant 2\sigma^2\end{cases}$$

实际计算时，λ_1 和 σ 可用其上界代替，λ_n 可用其下界（大于零）代替。

定理 2.11　假定矩阵 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$ 的对角线部分 $\boldsymbol{D}$ 为非奇，并且矩阵 $\boldsymbol{B}=\boldsymbol{D}^{-1}\cdot(\boldsymbol{L}+\boldsymbol{U})$ 满足如下条件：

（i） $\boldsymbol{B}$ 的所有元素均非负。

（ii） $\boldsymbol{B}$ 是不可约的，且 $0<S(\boldsymbol{B})<1$。

（iii） $\boldsymbol{B}$ 为对称矩阵。

则当逐次松弛法中松弛因子取为：

$$\omega_0=\frac{2}{1+\sqrt{1-S(\boldsymbol{B})^2}}$$

时，迭代矩阵 $\boldsymbol{L}_{\omega_0}$ 之谱半径应满足如下关系式：

$$\omega_0-1\leqslant S(\boldsymbol{L}_{\omega_0})<\sqrt{\omega_0-1}$$

并且仅当矩阵 $\boldsymbol{B}$ 为具有相容次序和性质"A"的矩阵时，上式左端的等号成立。

定理 2.11 说明对于满足其中三个条件的矩阵 $\boldsymbol{B}$，最优松弛因子的近似值可以按定理 2.9 中具有性质"A"的矩阵的公式进行计算，其相应渐近收敛速度的损失并不是很大的。

定理 2.11 的证明，可参阅[13]第 116 页，定理 2.10 的证明，可参阅[26]。

(3) 最优松弛因子的试算确定。

最后我们简单讨论一下松弛因子的试算确定问题。这一问题对于实际计算是很有意义的。一方面由于大多数矩阵目前还没有计算松弛因子最优值的理论公式，只能用试算确定。另一方面，即使有理论计算公式，其中往往有一些参数（例如 $S(\boldsymbol{B})$ 等），难于事先确定，通常也需要用试算的办法来估计它们，即真正求得 ω_{opt} 也是与试算分不开的。

最简单的试算确定办法，是从同一初始向量出发，取不同的松弛因子 ω 迭代相同次数（注意，迭代次数不应太少！），然后比较其相应剩余或误差，并选取使剩余或误差之模最小的松弛因子作为最优松弛因子的近似值。这一方法虽然简单，但往往是有效的，特别是当使用者需要求解多次具有相同系数矩阵的方程组时更是如此。另外，如果使用者对于所求解的方程组有较深入了解或者积累了一定经验，常常可以事先定出一个包含 ω_{opt} 的不大的区间

$[\omega_a, \omega_b]$，这样，就可以大大减少试算的次数，较快地确定 ω_{opt} 的近似值。自然，也可以采用优选法的原则来从区间 $[\omega_a, \omega_b]$ 中选取进行试算的 ω 值，以便更快地找到 ω_{opt} 的近似值。由于这些作法是明显易懂的，这里不再赘述。

另一个试算确定方法，是以知道最优松弛因子 ω_{opt} 的某些性质为前提的。例如，知道其理论计算公式(8.2.37)，又如，知道 ω_{opt} 恰为使 $\boldsymbol{L}_\omega$ 的按模最大特征值由实数变为复数（当 ω 由小增大时）的临界值等等。这时，可以将求解方程组的迭代过程与估算 ω_{opt} 的过程统一起来进行，使得估算 ω_{opt} 不致花费许多无用的计算时间，以达到节省计算量的目的。具体来说，可任选一个松弛因子 ω 来进行迭代，在迭代过程中，从逐次近似解间的增量 $\Delta\boldsymbol{u}^{(k)}$ 来判断该松弛因子是否需要修改，如果需要，则将其按某一办法修改之，然后继续迭代下去。这样反复修改多次后，就可找到一个与 ω_{opt} 较为接近的松弛因子，以后的迭代过程就用这个松弛因子计算下去，直到最终求得解答为止。修改松弛因子的原则，需视 ω_{opt} 的具体性质而定，例如，当 ω_{opt} 满足关系式(8.2.37)时，我们可以用 $\Delta\boldsymbol{u}^{(k)}$ 来估算 μ_1 之近似值，从而求得 ω_{opt}。由于 $\Delta\boldsymbol{u}^{(k)}=\boldsymbol{L}_\omega\cdot\Delta\boldsymbol{u}^{(k-1)}$，所以，当 k 相当大时，$\boldsymbol{L}_\omega$ 的按模最大特征值 $\lambda_1\approx\|\Delta\boldsymbol{u}^{(k)}\|/\|\Delta\boldsymbol{u}^{(k-1)}\|$（自然，这只在 $\omega<\omega_{opt}$ 时才成立）。这样，在迭代足够多次后，就可计算出 $\theta_k\approx\|\Delta\boldsymbol{u}^{(k)}\|/\|\Delta\boldsymbol{u}^{(k-1)}\|$，并用公式：

$$\mu_1\approx(\theta_k+\omega-1)/(\omega\sqrt{\theta_k})$$

来估计 μ_1，然后，再用公式(8.2.37)算出 ω_{opt} 的近似值 $\tilde{\omega}$。自然，还可以对 $\tilde{\omega}$ 重复上述过程，直到求得一个较好的松弛因子为止。另外，如果不知道形如(8.2.37)的关系式，仅知道 ω_{opt} 为 $\boldsymbol{L}_\omega$ 的按模最大特征值由实数变为复数的临界值时（例如，对于某些大地测量问题中的法方程式就是如此），我们可在迭代过程中检查 $\Delta\boldsymbol{u}^{(k)}$ 的变化情况来判断 $\boldsymbol{L}_\omega$ 的按模最大特征值的性质。若 $\Delta\boldsymbol{u}^{(k)}$ 之分量周期性地改变符号，则表明该特征值为复数，应将所用松弛因子减小一些。若 $\Delta\boldsymbol{u}^{(k)}$ 之分量按比例下降，则该特征值为实数，应将松弛因子增大。如此反复修改多次，最终即可估得一个较好的松弛因子。总之，根据要求解的问题中最优松弛因子的性质，利用迭代过程中的信息，将它估计出来，使得估算过程不致占用过多的计算时间是试算确定松弛因子的第二个方法的要点。读者可以根据这一精神灵活运用于实际问题中去。若对方法的细节感兴趣，可以参阅[19、20、21]，这里不再赘述。

应该指出，对于某些特殊问题（例如某些简单的椭圆型方程边值问题），矩阵 $\boldsymbol{B}$ 的按模最大特征值 μ_1 可以用某种办法事先估算出来。这时，即可按公式(8.2.37)算出 ω_{opt} 的近似值直接用于迭代过程中去。关于 μ_1 的一些事先的估算办法，可参阅[13、19]。

最后，还要指出，对于具有性质“A”的矩阵在求得 ω_{opt} 的近似值后，应该用比其稍微大一点的值作为实际计算的松弛因子。这是由于图 8.9 中所示曲线的性质和考虑到应使 $\|\boldsymbol{L}_\omega\|$ 尽可能小的缘故。实践经验亦表明这样作是有利的。

8.2.3 一阶线性定常迭代法的加速——切比雪夫半迭代法

对于任何一个一阶线性定常迭代法(8.2.7)，一般来说，我们都可以构造一个与之相关联的非定常迭代法，使其比(8.2.7)收敛得更快一些。这个非定常迭代法就是通常所谓的基于一阶线性定常迭代法(8.2.7)的切比雪夫半迭代法。

当(8.2.7)的迭代矩阵 G 的特征值均为实数时，切比雪夫半迭代法能够取得较好效果。但若 G 之特征值为复数，且不能包含在复平面上一个比较扁平的椭圆内时（如按模最大的

某几个特征值之虚部与实部相比较为相当大而又较分散时)，切比雪夫半迭代法是难于奏效的。我们将重点讨论 G 有实特征值之情形。

如果方程组的系数矩阵 $\boldsymbol{A}$ 对称正定，但不具有前节所述的那些性质(如性质"A"，非负性等)。用松弛法来求解时，可能收敛较慢。如果此时使用某种切比雪夫半迭代法，却可能取得较好效果，这一点希望读者注意。

下面来讨论构造切比雪夫半迭代法的一般原则和它的一种常用形式。

(一)一般原则

先从加速实数序列收敛的问题说起。通常，给定一个实数序列 $x_0, x_1, x_2, \cdots$，若其不收敛或收敛很慢，我们就可由它构造一个新的序列 $y_0, y_1, y_2, \cdots$，使得$\{y_i\}$是收敛的或是较$\{x_i\}$收敛更快。最简单的例子是：

$$y_0 = x_0$$

$$y_1 = \frac{x_0 + x_1}{2}$$

$$y_2 = \frac{x_0 + x_1 + x_2}{3}$$

$$y_3 = \frac{x_0 + x_1 + x_2 + x_3}{4}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

容易证明，若$\{x_i\}$收敛，则$\{y_i\}$亦收敛，并且，当$\{x_i\}$不收敛时，$\{y_i\}$亦可以是收敛的。

构造新序列$\{y_i\}$的更一般的办法，可按下列原则进行。考虑下列三角形系数表：

$$\begin{matrix} a_{00} & & & & \\ a_{10} & a_{11} & & & \\ a_{20} & a_{21} & a_{22} & & \\ a_{30} & a_{31} & a_{32} & a_{33} & \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{matrix}$$

其中

$$\sum_{k=0}^{m} a_{mk} = 1 \quad (m = 0, 1, 2, \cdots) \tag{8.2.42}$$

新序列$\{y_i\}$与序列$\{x_i\}$间的关系即可表为：

$$y_m = \sum_{k=0}^{m} a_{mk} x_k \quad (m = 0, 1, 2, \cdots)$$

即是说，我们是用原序列前 m 项的线性组合来得出新序列的第 m 项的。显然，当 $a_{mk}=0$ $(k<m)$ 及 $a_{mm}=1$ 时，$\{y_i\}$就是$\{x_i\}$。当 $a_{mk}=\dfrac{1}{m+1}$(所有 k)时，$\{y_i\}$ 就是前面例子中的序列。为了保证新序列与原序列有相同极限，通常应加以限制条件(8.2.42)。

完全类似，我们可将上面关于实数序列的原则应用到向量序列上来。考虑求解线性代数方程组 $\boldsymbol{Au}=\boldsymbol{b}$ 的一阶线性定常迭代法(8.2.7)，假设由其定义的向量序列 $\boldsymbol{u}^{(0)}, \boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}, \cdots$，收敛于方程组 $\boldsymbol{Au}=\boldsymbol{b}$ 的真解 $\boldsymbol{u}^*$，则可定义新的向量序列$\{\boldsymbol{v}^{(m)}\}$为：

$$\boldsymbol{v}^{(m)} = \sum_{k=0}^{m} a_{mk} \cdot \boldsymbol{u}^{(k)} \tag{8.2.43}$$

其中系数 a_{mk} 满足条件(8.2.42)。

(8.2.43)所定义的过程叫作关于一阶线性定常迭代法(8.2.7)的一种半迭代法。为了保

证向量序列$\{\boldsymbol{v}^{(m)}\}$也收敛至真解$\boldsymbol{u}^*$，必须对(8.2.43)中的系数α_{mk}加以限制条件(8.2.42)。因为若取$\boldsymbol{u}^{(0)}$为真解$\boldsymbol{u}^*$，则从(8.2.7)知$\boldsymbol{u}^{(m)}=\boldsymbol{u}^*$(对所有$m$)。此时，自然也要求

$$\boldsymbol{v}^{(m)}=\sum_{k=0}^{m}\alpha_{mk}\boldsymbol{u}^{(k)}=\Big(\sum_{k=0}^{m}\alpha_{mk}\Big)\boldsymbol{u}^*=\boldsymbol{u}^*$$

于是，推知条件(8.2.42)必须成立。

容易看出，在条件(8.2.42)的限制下，我们总可适当选取系数α_{mk}，使得序列$\{\boldsymbol{v}^{(m)}\}$收敛于方程组$\boldsymbol{A}\cdot\boldsymbol{u}=\boldsymbol{b}$的真解$\boldsymbol{u}^*$。为说明这一点，我们令：

$$\boldsymbol{\varepsilon}^{(m)}=\boldsymbol{u}^{(m)}-\boldsymbol{u}^*=\boldsymbol{G}^m\cdot\boldsymbol{\varepsilon}^{(0)}$$

$$\boldsymbol{\eta}^{(m)}=\boldsymbol{v}^{(m)}-\boldsymbol{u}^*$$

从(8.2.42)和(8.2.43)式，我们有：

$$\begin{aligned}\boldsymbol{\eta}^{(m)}&=\boldsymbol{v}^{(m)}-\boldsymbol{u}^*=\sum_{k=0}^{m}\alpha_{mk}\cdot\boldsymbol{u}^{(k)}-\boldsymbol{u}^*=\sum_{k=0}^{m}\alpha_{mk}\cdot\boldsymbol{u}^{(k)}-\boldsymbol{u}^*\sum_{k=0}^{m}\alpha_{mk}\\&=\sum_{k=0}^{m}\alpha_{mk}(\boldsymbol{u}^{(k)}-\boldsymbol{u}^*)=\sum_{k=0}^{m}\alpha_{mk}\cdot\boldsymbol{\varepsilon}^{(k)}=\Big(\sum_{k=0}^{m}\alpha_{mk}\cdot\boldsymbol{G}^k\Big)\boldsymbol{\varepsilon}^{(0)}\\&=P_m(\boldsymbol{G})\cdot\boldsymbol{\eta}^{(0)}\quad(\text{因为 }\boldsymbol{v}^{(0)}=\boldsymbol{u}^{(0)})\end{aligned}\tag{8.2.44}$$

其中

$$P_m(\boldsymbol{G})=\sum_{k=0}^{m}\alpha_{mk}\cdot\boldsymbol{G}^k$$

为使$\boldsymbol{\eta}^{(m)}\to0$，显然只需$\lim\limits_{m\to\infty}P_m(\boldsymbol{G})=0$。满足这一要求的多项式是很容易找到的，最简单的就是$P_m(x)=x^m$。此时$P_m(\boldsymbol{G})=\boldsymbol{G}^m$，由于$\boldsymbol{G}$之特征值之模小于1，自然就有$\lim\limits_{m\to\infty}P_m(\boldsymbol{G})=0$，即$\boldsymbol{v}^{(m)}\to\boldsymbol{u}^*$。究竟应该怎样选取多项式$P_m(x)$(注意！条件(8.2.42)现在变为$P_m(1)=1$，这是对多项式$P_m(x)$必须加上的限制条件)，才能使序列$\boldsymbol{v}^{(m)}$收敛最快？要解决上述问题，首先应该弄清楚非定常迭代法的收敛快慢如何衡量。

为说明简单，假定矩阵$\boldsymbol{G}$的特征向量$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$是线性无关的，其相应特征值为$\mu_1, \mu_2, \cdots, \mu_n$。于是，可将$\eta^{(0)}$表为：

$$\boldsymbol{\eta}^{(0)}=\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_n\boldsymbol{x}_n$$

从(8.2.44)便有：

$$\boldsymbol{\eta}^{(m)}=P_m(\mu_1)\alpha_1\boldsymbol{x}_1+P_m(\mu_2)\alpha_2\boldsymbol{x}_2+\cdots+P_m(\mu_n)\alpha_n\boldsymbol{x}_n$$

显然，$\boldsymbol{\eta}^{(m)}$收敛于零的速度由$\max\limits_{1<i<n}|P_m(\mu_i)|$决定。所以，为使迭代过程收敛最快，只需要选取$P_m(x)$为所有$m$阶多项式中使下式达极小值者即可：

$$S(P_m(\boldsymbol{G}))=\max_{1<i<n}|P_m(\mu_i)|=\min\tag{8.2.45}$$

$$P_m(1)=1$$

和8.2.2节中讨论渐近收敛速度时类似，定义如下数量为非定常迭代法(8.2.43)的渐近收敛速度：

$$R_\infty(P_m(\boldsymbol{G}))=\lim_{m\to\infty}\Big(-\frac{1}{m}\ln S(P_m(\boldsymbol{G}))\Big)\tag{8.2.46}$$

容易验证，当$P_m(x)=x^m$时，(8.2.46)与(8.2.23)是相同的，所以是其自然的推广。

由于μ_i事先不知道，所以极小化问题(8.2.45)不能求解。为实际应用方便，总是将其替之以如下极小化问题：

$$\begin{cases} [\max\limits_{\mu\in S_G}|P_m(\mu)|] = \min \quad (m\geqslant 0) \\ P_m(1) = 1 \end{cases} \tag{8.2.47}$$

其中 S_G 为包含矩阵 $\boldsymbol{G}$ 所有特征值 μ_i 的最小凸域。今后只讨论极小化问题(8.2.47)，并以其解答 $P_m(x)$ 作为要求的最优多项式。自然，我们有：

$$\max_{1\leqslant i\leqslant n}|P_m(\mu_i)|^{1/m}\leqslant \max_{\mu\in S_G}|P_m(\mu)|^{1/m} \tag{8.2.48}$$

(二)切比雪夫半迭代法

假定矩阵 $\boldsymbol{G}$ 的特征值 μ_i 均是实数，并满足如下不等式：

$$\alpha\leqslant\mu_i\leqslant\beta<1 \quad (\alpha<\beta) \tag{8.2.49}$$

这里，α 可以小于 -1。我们将指出，在这种情况下，用切比雪夫多项式构成的半迭代法，将使原来迭代过程得到加速。通常就称这样的半迭代法为切比雪夫半迭代法。

首先引入新变数 γ

$$\gamma=\gamma(\mu)=\frac{2\mu-(\alpha+\beta)}{\beta-\alpha} \tag{8.2.50}$$

显然，有：$\gamma(\alpha)=-1$，$\gamma(\beta)=1$，且当 μ 满足(8.2.49)时，$-1\leqslant\gamma\leqslant 1$，同时，也很容易验证：

$$z=\gamma(1)=\frac{2-(\alpha+\beta)}{\beta-\alpha}>1 \tag{8.2.51}$$

将新变数 γ 代入多项式 $P_m(\mu)$ 中，便得到另一个 γ 的多项式 $Q_m(\gamma)$：

$$Q_m(\gamma)=P_m\left(\frac{(\beta-\alpha)\gamma+\beta+\alpha}{2}\right)=P_m(\mu)$$

这样一来，便有：

$$\max_{\alpha\leqslant\mu\leqslant\beta}|P_m(\mu)|=\max_{-1\leqslant\gamma\leqslant 1}|Q_m(\gamma)| \tag{8.2.52}$$

于是，极小化问题(8.2.47)将变为：

$$\begin{cases} \max\limits_{-1\leqslant\gamma\leqslant 1}|Q_m(\gamma)| = \min \quad (m\geqslant 0) \\ Q_m(z)=1 \end{cases} \tag{8.2.53}$$

这个问题的解答是：

$$Q_m(\gamma)=T_m(\gamma)/T_m(z) \tag{8.2.54}$$

其中，$T_m(\gamma)$ 就是 m 阶切比雪夫多项式，其表达式为：

$$T_m(x)=\begin{cases} \cos(m\cdot\operatorname{arc}\cos x) & |x|\leqslant 1 \\ (-1)^m\cdot\cosh(m\cdot\operatorname{arc}\cosh|x|) & |x|\geqslant 1 \end{cases} \quad (m\geqslant 0)$$

(这一事实的详细证明，可参看[1]第 49 页或[19]pp. 302)。所以有：

$$P_m(\mu)=Q_m(\gamma)=Q_m\left(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha}\right)=\frac{T_m\left(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha}\right)}{T_m\left(\frac{2-(\beta+\alpha)}{\beta-\alpha}\right)}$$

此外，

$$\min\left(\max_{\alpha\leqslant\mu\leqslant\beta}|P_m(\mu)|\right)=\max_{\alpha\leqslant\mu\leqslant\beta}\left|T_m\left(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha}\right)\Big/T_m(z)\right|=\frac{1}{T_m(z)} \tag{8.2.55}$$

实际计算 $\boldsymbol{v}^{(m)}$ 时，将采用切比雪夫多项式的如下三项递推关系式：

$$\begin{cases} T_0(x)=1, \quad T_1(x)=x \\ T_{m+1}(x)=2xT_m(x)-T_{m-1}(x) \end{cases} \quad (m\geqslant 1) \tag{8.2.56}$$

于是

$$\begin{aligned}\boldsymbol{\eta}^{(m+1)}&=T_{m+1}\left(\frac{2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}}{\beta-\alpha}\right)\cdot T_{m+1}(z)^{-1}\cdot\boldsymbol{\varepsilon}^{(0)}\\&=\left\{\left[2\left(\frac{2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}}{\beta-\alpha}\right)\cdot T_m\left(\frac{2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}}{\beta-\alpha}\right)\right.\right.\\&\qquad\left.\left.-T_{m-1}\left(\frac{2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}}{\beta-\alpha}\right)\right]\Big/T_{m+1}(z)\right\}\cdot\boldsymbol{\varepsilon}^{(0)}\\&=2\left[\frac{2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}}{\beta-\alpha}\right]\frac{T_m(z)}{T_{m+1}(z)}\boldsymbol{\eta}^{(m)}-\frac{T_{m-1}(z)}{T_{m+1}(z)}\boldsymbol{\eta}^{(m-1)}\end{aligned}$$

将 $\boldsymbol{\eta}^{(m)}=\boldsymbol{v}^{(m)}-\boldsymbol{u}^*$ 代入上式，并注意到从(8.2.7)有：

$$2\left[z\cdot\boldsymbol{I}-\frac{2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}}{\beta-\alpha}\right]\frac{T_m(z)}{T_{m+1}(z)}\cdot\boldsymbol{u}^*=\frac{4}{\beta-\alpha}\,\frac{T_m(z)}{T_{m+1}(z)}\boldsymbol{d}$$

并利用(8.2.56)式便可得

$$\begin{cases}\boldsymbol{v}^{(m+1)}=2\left[\dfrac{2}{\beta-\alpha}\boldsymbol{G}-\dfrac{\beta+\alpha}{\beta-\alpha}\boldsymbol{I}\right]\dfrac{T_m(z)}{T_{m+1}(z)}\boldsymbol{v}^{(m)}-\dfrac{T_{m-1}(z)}{T_{m+1}(z)}\boldsymbol{v}^{(m-1)}+\dfrac{4}{\beta-\alpha}\,\dfrac{T_m(z)}{T_{m+1}(z)}\boldsymbol{d}\\ \boldsymbol{v}^{(1)}=\dfrac{2}{2-(\beta+\alpha)}(\boldsymbol{G}\boldsymbol{v}^{(0)}+\boldsymbol{d})-\dfrac{\beta+\alpha}{2-(\beta+\alpha)}\boldsymbol{v}^{(0)}\end{cases}$$

若令 $\rho_1=1$, $\rho_m=2zT_{m-1}(z)/T_m(z)$，从递推关系式(8.2.56)容易证明 ρ_i 有如下递推关系式：

$$\begin{cases}\rho_1=1, \quad \rho_2=\dfrac{2z^2}{2z^2-1}\\ \rho_{m+1}=\left(1-\dfrac{1}{4z^2}\rho_m\right)^{-1} \quad (m=2,\ 3,\ \cdots)\end{cases} \tag{8.2.57}$$

前面的计算 $\boldsymbol{v}^{(m+1)}$ 的公式现在即可写为：

$$\boldsymbol{v}^{(m+1)}=\frac{\rho_{m+1}}{2-(\alpha+\beta)}\{[2\boldsymbol{G}-(\beta+\alpha)\boldsymbol{I}]\boldsymbol{v}^{(m)}+2\boldsymbol{d}\}+(1-\rho_{m+1})\boldsymbol{v}^{(m-1)} \tag{8.2.58}$$

(8.2.57)和(8.2.58)两式就是对于一阶线性定常迭代法(8.2.7)的切比雪夫半迭代法计算公式。

现在来说明切比雪夫半迭代法将使原来的迭代过程加速收敛。从(8.2.45)、(8.2.48)及(8.2.55)知道：

$$S(P_m(\boldsymbol{G}))\leqslant\frac{1}{T_m(z)}$$

但是

$$T_m(z)=\frac{1}{2}(e^{m\cdot\cosh^{-1}z}+e^{-m\cdot\cosh^{-1}z})$$

(因为 $z>1$)，利用关系式

$$\cosh^{-1}z=\ln(z+\sqrt{z^2-1})$$

立即可以推得：

$$(T_m(z))^{-1}=\frac{2\tau^m}{1+\tau^{2m}}$$

其中
$$\tau=\frac{\sigma}{1+\sqrt{1-\sigma^2}}<1;\quad \sigma=\frac{1}{z}$$

从半迭代法渐近收敛速度的定义(8.2.46)得知:

$$R_\infty(P_m(\boldsymbol{G}))\geqslant\lim_{m\to\infty}\left(-\frac{1}{m}\lg\frac{1}{T_m(z)}\right)=\lim_{m\to\infty}\left(-\frac{1}{m}\lg\frac{2\tau^m}{1+\tau^{2m}}\right)=-\lg\tau$$

另外有:

$$R_\infty(\boldsymbol{G})\leqslant R_\infty(P_1(\boldsymbol{G}))\leqslant-\lg\frac{2\tau}{1+\tau^2}=-\lg\sigma$$

这样一来,便有:

$$\frac{R_\infty(P_m(\boldsymbol{G}))}{(R_\infty(\boldsymbol{G}))^{1/2}}\leqslant\frac{-\lg\tau}{(-\lg\sigma)^{1/2}}\to\sqrt{2}^{㊀}\quad(\text{当 }\sigma\to1-)$$

也就是说,当σ接近于1时,切比雪夫半迭代法的渐近收敛速度将较原来一阶线性定常迭代法(8.2.7)的渐近收敛速度大得多。这样,当$\boldsymbol{G}$的特征值都是实数时,采用切比雪夫半迭代法来加速原有的一阶线性定常迭代法,总是会得到收敛速度的改进的。

当$\boldsymbol{G}$的特征值为复数时切比雪夫半迭代法的加速效果将大为减小。如果包含矩阵$\boldsymbol{G}$全部特征值的最小凸集S_G与某个较扁平的椭圆相近(例如$\boldsymbol{G}$的特征值的虚部均较小),可以证明,适当选取参数α及β后,前述的切比雪夫半迭代法仍然是收敛的,但加速效果已大为降低。随着该椭圆扁平程度的减少,加速的效果愈降低,直至该椭圆变为某个圆$|z|\leqslant R$时,根据函数逼近论的熟知结果,在复平面的圆$|z|\leqslant R$上,与零偏差最小的m次多项式为z^m。即可看到,前述的加速迭代过程的办法将是无济于事的,此时最有效的是将原来的迭代过程本身重复m次。

(三)简单迭代法的切比雪夫加速

如前节所述,切比雪夫半迭代法仅对矩阵$\boldsymbol{G}$的特征值全为实数时才是有效的。所讨论过的几种一阶线性定常迭代法中,逐次松弛法的迭代矩阵当$\omega\sim\omega_{opt}$时,一般具有复特征值,故对其使用切比雪夫加速收益是不大的。

虽然逐次松弛法的一种变形——对称松弛法的相应特征值为实数,经切比雪夫加速后其收敛速度较高,但由于其较为复杂,故实际上较少使用。这里,仅讨论基于简单迭代法的切比雪夫半迭代法。

假定方程组$\boldsymbol{Au}=\boldsymbol{b}$的系数矩阵是对称正定的。其相应的简单迭代法可以表为:

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{B}\boldsymbol{u}^{(m)}+\boldsymbol{c}$$

其中$\boldsymbol{B}=\boldsymbol{I}-\boldsymbol{D}^{-1}\cdot\boldsymbol{A}$, $\boldsymbol{D}=\mathrm{diag}(a_{ii})$为正定的对角线矩阵,$\boldsymbol{c}=\boldsymbol{D}^{-1}\cdot\boldsymbol{b}$。

由于$\widetilde{\boldsymbol{B}}=\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{B}\boldsymbol{D}^{-\frac{1}{2}}=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$为对称矩阵,所以,$\widetilde{\boldsymbol{B}}$的特征值为实数,从而$\boldsymbol{B}$的特征值均为实数。同时,由于$\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$是正定矩阵,$\boldsymbol{B}$的特征值还应该是小于1的实数。这样一来,便可以找到常数α和β,使得$\boldsymbol{B}$的特征值μ_i满足关系式(8.2.49)。根据(8.2.57)和(8.2.58),采用简单迭代法的切比雪夫半迭代法就可表为:

$$\boldsymbol{v}^{(m+1)}=\frac{\rho_{m+1}}{2-(\beta+\alpha)}\{[2\boldsymbol{B}-(\beta+\alpha)\boldsymbol{I}]\boldsymbol{v}^{(m)}+2\boldsymbol{c}\}+(1-\rho_{m+1})\boldsymbol{v}^{(m-1)}\qquad(8.2.59)$$

$$(m=1,\ 2,\ \cdots)$$

其中ρ_m按(8.2.57)计算。

如果矩阵$\boldsymbol{A}$是形如(8.2.36)的具有相容次序和性质"A"的矩阵,从定理2.9得知矩阵

㊀ 对于这一极限关系的详细计算,可参阅[19] pp. 353。

$\boldsymbol{B}$ 的谱半径 $S(\boldsymbol{B})<1$，故可以令 $\beta=-\alpha=S(\boldsymbol{B})$。因此，切比雪夫半迭代法的计算公式可写为：

$$\begin{cases} \boldsymbol{v}^{(m+1)}=\rho_{m+1}(\boldsymbol{B}\boldsymbol{v}^{(m)}+\boldsymbol{c})+(1-\rho_{m+1})\cdot\boldsymbol{v}^{(m-1)} \quad (m=0,\ 1,\ 2,\ \cdots) \\ \rho_1=1 \\ \rho_2=2/(2-S(\boldsymbol{B})^2) \\ \rho_{m+1}=\left(1-\dfrac{1}{4}\cdot S(\boldsymbol{B})^2\cdot\rho_m\right)^{-1} \quad (m=2,\ 3,\ \cdots) \end{cases} \tag{8.2.60}$$

根据前节的讨论，其渐近收敛速度应为：

$$R_\infty(P_m(\boldsymbol{B}))\geqslant-\lg\frac{S(\boldsymbol{B})}{1+\sqrt{1-S(\boldsymbol{B})^2}}=-\frac{1}{2}\lg\left(\frac{1-\sqrt{1-S(\boldsymbol{B})^2}}{1+\sqrt{1-S(\boldsymbol{B})^2}}\right)$$

但从(8.2.41)得知，此时逐次松弛法($\omega=\omega_{opt}$ 时)的渐近收敛速度为：

$$R_\infty(\boldsymbol{L}\omega_{opt})=-\lg(\omega_{opt}-1)=-\lg\left(\frac{1-\sqrt{1-S(\boldsymbol{B})^2}}{1+\sqrt{1-S(\boldsymbol{B})^2}}\right).$$

所以，当矩阵 $\boldsymbol{A}$ 为对称正定且具有性质“A”和相容次序的矩阵时，从渐近收敛速度来说，逐次松弛法($\omega=\omega_{opt}$)为基于简单迭代法的切比雪夫半迭代法的两倍。

我们也可以根据二循环矩阵的特点，将切比雪夫半迭代法的公式加以改写，使其收敛速度增加一倍，从而达到与逐次松弛法相同的效果(详细讨论可参见[13] pp. 138)。不过，其使用起来不够方便，故这里不再讨论。总之，当系数矩阵 $\boldsymbol{A}$ 为对称正定矩阵且具有性质“A”和相容次序时，采用逐次超松弛法一般说来是较为有利的。

如果矩阵 $\boldsymbol{A}$ 是对称正定矩阵，但不具有性质“A”或非负性等特点，则一般来说，逐次松弛法将收敛较慢。此时，采用切比雪夫半迭代法可能获得较好的效果。但按照(8.2.59)计算时，其存储需要量将有所增加(需要存放相邻两次的近似向量 $\boldsymbol{v}^{(m)}$、$\boldsymbol{v}^{(m-1)}$)。也可以导出只需要一片存储区的切比雪夫半迭代法公式，其收敛速度仅稍微降低一些(详见本书第十三章)。可视具体情况来决定迭代方法的选取。

8.2.4 分块迭代法

(一)分块迭代法的计算过程

前面讨论的迭代法通常称之为显式(或点)迭代法。其特点是求新的近似解时逐个分量地决定，即是说，每次从已有的近似解分量求出一个新的近似解分量，如此逐个地求下去，直到求得全部分量为止。然后，重复此过程。由于每次只求一个分量，故可直接显式地求出其值，所以称为显式迭代法。又由于一个分量通常与椭圆型偏微分方程边值问题数值求解中一个网格点上的未知数相对应，所以又称为点迭代法。

这里要讨论的分块迭代法在确定新的近似分量时是分组确定的。即先将要求解方程组的所有方程式和未知数进行分组，使每一方程式和未知数均属于且仅属于一个组。进行迭代时，从每一个方程式组中同时确定相应的一组新的近似解分量，如此按组计算下去，直到求得全部新的近似解分量为止。然后重复此过程。从矩阵形式来看，就是先将系数矩阵 $\boldsymbol{A}$，解向量 $\boldsymbol{u}$ 和自由项 $\boldsymbol{b}$ 进行分块，然后将每一子块视为一个元素，并按照点迭代法的类似公式进行迭代。例如，将(8.2.1)的系数矩阵，解向量和自由项进行如下分块：

$$\begin{bmatrix} A_{11}, & A_{12}, & \cdots, & A_{1M} \\ A_{21}, & A_{22}, & \cdots, & A_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ A_{M1}, & A_{M2}, & \cdots, & A_{MM} \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_M \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_M \end{bmatrix} \tag{8.2.61}$$

其中，A_{ii} 为 n_i 阶方阵；U_i、B_i 为 n_i 维向量；A_{ij} 为 $n_i\times n_j$ 维矩阵。然后，将 A_{ij} 及 U_i、B_i 等视为元素。仿照(8.2.5)式，即可得到分块简单迭代法的下列计算公式：

$$\begin{cases} U_1^{(k+1)} = -A_{11}^{-1}\cdot A_{12}U_2^{(k)} + \cdots - A_{11}^{-1}\cdot A_{1M}U_M^{(k)} + A_{11}^{-1}\cdot B_1 \\ U_2^{(k+1)} = -A_{22}^{-1}A_{21}U_1^{(k)} - A_{22}^{-1}\cdot A_{23}U_3^{(k)}\cdots - A_{22}^{-1}A_{2M}U_M^{(k)} + A_{22}^{-1}\cdot B_2 \\ \cdots\cdots\cdots\cdots \\ U_M^{(k+1)} = -A_{MM}^{-1}\cdot A_{M1}U_1^{(k)} - A_{MM}^{-1}A_{M2}\cdot U_2^{(k)} - \cdots - A_{MM}^{-1}\cdot A_{MM-1}U_{M-1}^{(k)} + A_{MM}^{-1}\cdot B_M \end{cases}$$

或者写为：

$$\begin{cases} A_{11}U_1^{(k+1)} = \qquad\qquad -A_{12}U_2^{(k)} - \cdots - A_{1M}U_M^{(k)} + B_1 \\ A_{22}U_2^{(k+1)} = -A_{21}U_1^{(k)} \qquad\qquad -A_{23}U_3^{(k)} - \cdots - A_{2M}U_M^{(k)} + B_2 \\ \cdots\cdots\cdots\cdots \\ A_{MM}U_M^{(k+1)} = -A_{M1}U_1^{(k)} - A_{M2}U_2^{(k)} - \cdots - A_{MM-1}U_{M-1}^{(k)} + B_M \end{cases} \tag{8.2.62}$$

进行迭代时，可以从第一组方程式解出 n_1 个未知数 $U_1^{(k+1)}$，然后从第二组方程式解出 n_2 个未知数 $U_2^{(k+1)}$，如此等等。

由于从每一个组中解出相应未知数需要求解某些阶数较低的方程组 $A_{ii}U_i^{(k+1)}=v$，不能象点迭代那样将未知数显式地解出来，所以这种迭代法称为隐式迭代法。此外，由于矩阵的子块或解向量的子向量与点迭代法中矩阵元素或向量分量的位置相当，故又称为块迭代法。在解边值问题时，常常把一条或几条网格线上的所有未知数分在一个组内，故又称为线迭代法。

可以用直接法求解块迭代过程中的低阶方程组。由于这些子方程组阶数较低且往往具有某种特殊形状，例如三对角线型等等，故其求解是容易的。这就使得块迭代法有时较为有效。

(二)块松弛法

现在来讨论一种比较常用的分块迭代法——块松弛法。

假设方程组 $Au=b$ 的系数矩阵 A，解向量 u 及自由项 b，有形如(8.2.61)的分块形式。如果我们将子块 A_{ij}、U_j、B_j 视为元素，完全仿照逐次松弛法(8.2.26)的计算公式，立即可以得出如下块松弛法的计算公式：

$$A_{ii}\cdot U_i^{(k+1)} = (1-\omega)A_{ii}U_i^{(k)} + \omega\left\{B_i - \sum_{j=1}^{i-1}A_{ij}U_j^{(k+1)} - \sum_{j=i+1}^{M}A_{ij}U_j^{(k)}\right\} \tag{8.2.63}$$

$$(i=1,\ 2,\ \cdots,\ M)$$

其中 ω 为松弛因子。

求解 $U_i^{(k+1)}$ 时，可以采用高斯消去法。许多情况下，A_{ii} 为三对角线型或带型矩阵，这时可以采用本章§8.1所叙述的追赶法或带型矩阵消去法来求解。当矩阵 A 是对称正定矩阵时，A_{ii} 亦应是对称正定的，追赶法和消去法的数值稳定性均能得到保证，其工作量亦比较节省(参看本章§8.1)。

如果 A_{ii} 为对称正定的三对角线型矩阵，还可先将 A_{ii} 分解为如下形式：

$$A_{ii} = DT^{T}TD \tag{8.2.64}$$

其中，D 为正定对角线型矩阵；T 为单位上三角形矩阵(仅有一条次对角线元非零)。反复求解 $U_i^{(k+1)}$ 时，即变为求解方程组：

$$T^T T(DU_i^{(k+1)}) = D^{-1}v \tag{8.2.65}$$

其所需完成的总运算量将更加减少(详见[19] pp. 442)。类似作法，也可应用到 A_{ii} 为五对角线或多对角线的情形。

块松弛法的收敛条件与逐次点松弛法类似。例如，完全仿照证明定理 2.6 的办法，可以证明如下定理：

定理 2.12　若 A 为对称矩阵，A_{ii} 为正定矩阵($i=1, 2, \cdots, M$)，且 $0<\omega<2$，则逐次块松弛法(8.2.63)收敛的充要条件为 A 是正定矩阵。

逐次点松弛法中的"性质 A"亦可推广到逐次块松弛法中来，并有与定理 2.9 类似的结果等等。例如，对应于(8.2.61)中矩阵 A 的分块形式，我们可定义如下 M 阶矩阵 z：

$$z=[z_{ij}] \quad z_{ij}=\begin{cases} 0, & 若\ A_{ij}=0 \\ 1, & 若\ A_{ij}\neq 0 \end{cases} \tag{8.2.66}$$

当矩阵 z 具有"性质 A"时，我们就称矩阵 A 有"性质 $A^{(\pi)}$"。当矩阵 z 有相容次序时，矩阵 A 就称为具有 π 相容次序(我们用符号 π 表示该种确定的分块方式)。对于有性质 $A^{(\pi)}$ 的 π 相容次序对称正定矩阵 A，逐次块松弛法(8.2.63)的最优松弛因子为：

$$\omega_{opt}^{(\pi)} = \frac{2}{1+\sqrt{1-\bar{\mu}_{(\pi)}^2}} \tag{8.2.67}$$

相应迭代矩阵 $L_{\omega_{opt}^{(\pi)}}$ 的谱半径为：

$$S(L_{\omega_{opt}^{(\pi)}}) = \omega_{opt}^{(\pi)} - 1 \tag{8.2.68}$$

其中，$\bar{\mu}_{(\pi)}$ 为矩阵 $B=I-D^{-1}\cdot A$ 的谱半径；$D=\mathrm{diag}(A_{ii})$。关于这方面的详细讨论，可以参见[19] pp. 445。

逐次块松弛法的实际效果有时是较显著的。特别是采取将 A_{ii} 分解为 DT^TTD 的措施后，每迭代一次的工作量与点迭代相差不多，但收敛速度却提高了，因而可以获得较好效果。此外，某些矩阵没有性质 A，但适当分块后却具有性质 $A^{(\pi)}$，例如，拉普拉斯方程的九点差分方程式以及重调和差分方程的相应矩阵就是如此。这样一来，对这些问题，就可以应用逐次块松弛法求解它们，从而获得比点松弛法较高的收敛速度。还有一些情况，例如边值问题的求解区域为狭长的带状或是对角线上子块 A_{ii} 的模比较大时，使用逐次块松弛法也是较为有利的。所以，根据问题的具体情况，适当使用逐次块松弛法，有时可以获得较好效果。但是，应该指出，由于每次迭代的计算工作量增加，尽管收敛速度有所提高，但有时却得不到减少总的计算时间的效果，特别是当矩阵 A_{ii} 的形状较复杂，或者求解它们的程序效率不高时，更是如此。

除块松弛法外，还有一种对某些特殊问题有效的方法，即所谓隐式交替方向迭代法。它也可以看作是一种分块迭代法，但其分块的方式不是固定不变的，而是采取两种分块方式交替使用的办法，并且在分块时引入了某些加速收敛的参数，故与一般的块迭代法又有所不同。

交替方向法对于满足其可交换条件的某些问题(例如，矩形区域上波松方程第一边值问题的相应差分方程)是非常有效的。此外，实际计算的经验表明，即使可交换条件不满足，交替方向法有时还是可以有效地使用。例如，对于石油工业和核反应堆物理中的某些问题，交

替方向法就是有效方法之一。不过,对于大多数不满足可交换条件的情况,交替方向法如何有效使用的问题还未得到解决。由于这一原因,交替方向法的使用范围较为狭窄,这里不再讨论。读者可以参阅[7、13、14、19]。

8.2.5 共轭斜量法

共轭斜量法是一种非线性迭代法,它适用于系数矩阵为对称正定的情况。由于不需要选取任何迭代参数,所以使用比较方便。同时,如果没有舍入误差,理论上它能保证最多迭代 n 步(n 为方程组的阶数)便求得精确解,因而实质上也是一种直接法。目前,共轭斜量法的使用已较为普遍。实际计算表明,与前面讨论过的几种迭代法相较,在很多情况下其优越性是显著的。此外,共轭斜量法还可以直接推广至求解非线性方程组和求函数极小值问题,并已成为目前求解非线性方程组和函数极小值的有效方法之一。共轭斜量法的主要缺点是与其它迭代法相较,需要的存储量较大(大约需要 $3n$ 至 $4n$ 个单元来存放近似解及余量等),此外,每迭代一步需要的计算量也大一些。但是,它和所有迭代法一样,对于高阶稀疏矩阵问题都不需要 n^2 个单元来存放矩阵的元素。所以,对于这类问题来说,只要计算机的存储容量不是很小,上述存储量方面的缺点不是很重要的。

(一)等价极小值问题

如果我们要求解线性方程组:

$$\boldsymbol{Au}=\boldsymbol{b} \quad 或 \quad \sum_{k=1}^{n} a_{ik}u_k=b_i \quad (i=1,\ 2,\ \cdots,\ n) \tag{8.2.1}$$

并且在本节中我们假定其中系数矩阵 $\boldsymbol{A}$ 为对称正定矩阵。

这一问题可以转化为寻找下列二次函数的极小值问题:

$$F(\boldsymbol{u})=\frac{1}{2}(\boldsymbol{Au},\ \boldsymbol{u})-(\boldsymbol{b},\ \boldsymbol{u})=\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}u_iu_k-\sum_{i=1}^{n} b_iu_i \tag{8.2.69}$$

上述事实的证明是很简单的。因为利用矩阵 $\boldsymbol{A}$ 的对称性,不难验证二次函数 $F(\boldsymbol{u})$ 对 u_i 的偏导数为:

$$\frac{\partial F}{\partial u_i}=\sum_{k=1}^{n} a_{ik}u_k-b_i=-r_i$$

即是说:

$$\operatorname{grad} F(\boldsymbol{u})=\boldsymbol{Au}-\boldsymbol{b}=-\boldsymbol{r} \tag{8.2.70}$$

所以,将任意向量 $\boldsymbol{u}$ 代入方程式(8.2.1)中所得的剩余向量 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{Au}$ 与函数 F 在该向量处的梯度向量反号。当函数 F 在某个向量处取极小值时,自然其梯度应为零向量,所以,与 $\boldsymbol{u}$ 相应的剩余向量 $\boldsymbol{r}$ 亦为零向量,即 $\boldsymbol{u}$ 为方程式(8.2.1)的解答。另一方面,如果 $\boldsymbol{u}$ 为方程式(8.2.1)的解答,那么函数 F 在向量 $\boldsymbol{u}$ 处之梯度应为零向量,所以,$\boldsymbol{u}$ 是函数 F 的稳定值。但是,因为系数矩阵 $\boldsymbol{A}$ 是正定矩阵,故使其梯度为零的向量只有一个;并且用泰勒展开的办法也容易验证函数 F 在该处只取极小值。这样一来,我们就有如下定理:

定理 2.13 任意向量 $\boldsymbol{u}$ 为对称正定方程组(8.2.1)的解的充要条件是,向量 $\boldsymbol{u}$ 使(8.2.69)所定义的二次函数 $F(\boldsymbol{u})$ 达到极小值。

根据这个定理,我们就可以把求解对称正定线性代数方程组(8.2.1)的问题转化为求函数 $F(\boldsymbol{u})$ 的极小值问题。从这个思想出发建立起来的迭代法,一般称之为极小化方法,共轭斜量法就是其中之一。

(二)共轭斜量法的基本步骤及计算公式

各种极小化方法,例如,最速下降法、共轭方向法、共轭斜量法以及对称正定矩阵的松弛法等等,其计算的基本步骤大都是从任意初始解向量 $\boldsymbol{u}$ 出发,沿着某个适当的方向向量 $\boldsymbol{p}$ 来修正它,从而得到一个新的近似解向量 $\boldsymbol{u}'=\boldsymbol{u}+t\boldsymbol{p}$($t$ 为参数),并使得函数 F 在 $\boldsymbol{u}'$ 处的值小于在 $\boldsymbol{u}$ 处的值。然后对 $\boldsymbol{u}'$ 再重复类似的步骤。这样不断地对其修正下去,每步均使函数 F 之值减少,最终即可求得使函数 $F(\boldsymbol{u})$ 达极小值的解向量 $\boldsymbol{u}^*$。关键问题在于每一步的修正方向 $\boldsymbol{p}$ 如何选择,各种极小化方法之间的区别也在于此。为了说明共轭斜量法中修正方向 $\boldsymbol{p}$ 的选择问题,必须对函数 $F(\boldsymbol{u})$ 的值在修正前后的变化有进一步的了解。

很明显,我们有如下等式:

$$\begin{aligned}F(\boldsymbol{u}')&=F(\boldsymbol{u}+t\boldsymbol{p})=\frac{1}{2}[\boldsymbol{A}(\boldsymbol{u}+t\boldsymbol{p}),\ (\boldsymbol{u}+t\boldsymbol{p})]-(\boldsymbol{b},\ \boldsymbol{u}+t\boldsymbol{p})\\&=\frac{1}{2}(\boldsymbol{A}\boldsymbol{u},\ \boldsymbol{u})+t(\boldsymbol{A}\boldsymbol{u},\ \boldsymbol{p})+\frac{1}{2}t^2(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})-(\boldsymbol{b},\ \boldsymbol{u})-t(\boldsymbol{b},\ \boldsymbol{p})\\&=\frac{1}{2}t^2(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})-t(\boldsymbol{r},\ \boldsymbol{p})+F(\boldsymbol{u})\end{aligned}\tag{8.2.71}$$

自然,决定参数 t 的办法是使 $F(\boldsymbol{u}')$ 尽可能地小,这样便有如下条件:

$$\frac{dF(\boldsymbol{u}')}{dt}=t(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})-(\boldsymbol{r},\ \boldsymbol{p})=0$$

由此得知:

$$t_{\min}=\frac{(\boldsymbol{r},\ \boldsymbol{p})}{(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})},\ \text{其中,}\ \boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}\tag{8.2.72}$$

(注意,由于 $d^2F/dt^2=(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})>0$,所以,由上式决定的 $t_{\min}$ 必使 F 达到极小值)。

将这个 t 值代入 $F(\boldsymbol{u}')$ 的表达式(8.2.71)中,很容易求得:

$$\Delta F=F(\boldsymbol{u}+t_{\min}\cdot\boldsymbol{p})-F(\boldsymbol{u})=-\frac{1}{2}\frac{(\boldsymbol{r},\ \boldsymbol{p})^2}{(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})}<0\quad(\text{当}(\boldsymbol{r},\ \boldsymbol{p})\neq 0)\tag{8.2.73}$$

于是,当 $(\boldsymbol{r},\ \boldsymbol{p})\neq 0$ 时,$F(\boldsymbol{u}')$ 之值总是小于 $F(\boldsymbol{u})$ 的。自然,修正方向 $\boldsymbol{p}$ 不能与被修正向量 $\boldsymbol{u}$ 相应的余量值 $\boldsymbol{r}$ 正交,否则 $\Delta F=0$,便得不到任何改进。但是,可以证明,相应于修正后向量 $\boldsymbol{u}'$ 的余量 $\boldsymbol{r}'=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}'$,必与修正方向 $\boldsymbol{p}$ 正交。这是因为:

$$\begin{aligned}(\boldsymbol{r}',\ \boldsymbol{p})&=(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}',\ \boldsymbol{p})=(\boldsymbol{b}-(\boldsymbol{A}\boldsymbol{u}+t_{\min}\cdot\boldsymbol{A}\boldsymbol{p}),\ \boldsymbol{p})\\&=(\boldsymbol{r},\ \boldsymbol{p})-t_{\min}(\boldsymbol{A}\boldsymbol{p},\ \boldsymbol{p})=0\end{aligned}$$

从上述简单讨论得知,只要修正方向 $\boldsymbol{p}$ 不与被修正的向量 $\boldsymbol{u}$ 相应的余量 $\boldsymbol{r}$ 正交,我们总可以使函数 F 之值减小,从而得到一个较好的近似解 $\boldsymbol{u}'$。人们自然会想到,函数 F 之值增加最快的方向是其梯度方向,那么,取函数 F 在向量 $\boldsymbol{u}$ 处的负梯度方向(即剩余向量)作为修正方向 $\boldsymbol{p}$ 似乎是最好的。按照这个思想导出的方法就是所谓的最速下降法,其计算公式很容易从前面的讨论和(8.2.72)式推得如下:

任取初始向量 $\boldsymbol{u}^{(0)}$

$$\begin{cases}\boldsymbol{r}^{(k)}=\boldsymbol{b}-\boldsymbol{A}\cdot\boldsymbol{u}^{(k)}\\ t_k=\dfrac{(\boldsymbol{r}^{(k)},\ \boldsymbol{r}^{(k)})}{(\boldsymbol{A}\boldsymbol{r}^{(k)},\ \boldsymbol{r}^{(k)})}\quad(k=0,\ 1,\ 2,\ \cdots)\\ \boldsymbol{u}^{(k+1)}=\boldsymbol{u}^{(k)}+t_k\cdot\boldsymbol{r}^{(k)}\end{cases}\tag{8.2.74}$$

从(8.2.73)可知,按最速下降法计算一步,函数 F 的值将变化 ΔF_k:

$$\Delta F_k=-\frac{1}{2}\frac{(\boldsymbol{r}^{(k)},\ \boldsymbol{r}^{(k)})}{(\boldsymbol{A}\boldsymbol{r}^{(k)},\ \boldsymbol{r}^{(k)})}<0\quad(\text{只要 } \boldsymbol{r}^{(k)}\neq 0)\tag{8.2.75}$$

由此可见，函数 F 之值将单调减少直至其极小值；$\boldsymbol{u}^{(k)}$ 亦将收敛至要求的解向量 $\boldsymbol{u}^*$。然而，理论分析与实际计算均表明最速下降法收敛较慢。特别是对于病态的方程组更是如此，因而，目前最速下降法使用得较少。

在共轭斜量法中，修正方向的选择是按照满足所谓共轭关系的原则来进行的。具体来说，从初始向量 $\boldsymbol{u}^{(0)}$ 出发，第一步仍取负梯度方向为修正方向，即取 $\boldsymbol{p}^{(1)}=\boldsymbol{r}^{(0)}=-\operatorname{grad}F(\boldsymbol{u}^{(0)})$。于是有：

$$\begin{cases}\boldsymbol{u}^{(1)}=\boldsymbol{u}^{(0)}+q_1\cdot\boldsymbol{p}^{(1)}=\boldsymbol{u}^{(0)}+q_1\cdot\boldsymbol{r}^{(0)}\\ q_1=\dfrac{(\boldsymbol{r}^{(0)},\ \boldsymbol{r}^{(0)})}{(\boldsymbol{A}\boldsymbol{r}^{(0)},\ \boldsymbol{r}^{(0)})}=\dfrac{(\boldsymbol{r}^{(0)},\ \boldsymbol{p}^{(1)})}{(\boldsymbol{A}\boldsymbol{p}^{(1)},\ \boldsymbol{p}^{(1)})}\end{cases}\tag{8.2.76}$$

对于以后各步，例如第 k 步$(k\geqslant 2)$，修正方向不再取为 $\boldsymbol{r}^{(k-1)}$，而是在通过点 $\boldsymbol{u}^{(k-1)}$ 并由向量 $\boldsymbol{r}^{(k-1)}$ 和 $k-1$ 步的修正方向 $\boldsymbol{p}^{(k-1)}$ 所作的二维平面 π_k 内，找出使函数 F 减少最快的方向作为修正方向 $\boldsymbol{p}^{(k)}$。如图 8.10 所示，通过点 $\boldsymbol{u}^{(k-1)}$ 并由向量 $\boldsymbol{r}^{(k-1)}$ 和 $\boldsymbol{p}^{(k-1)}$ 所作的二维平面 π_k，与 n 维椭球面 $F(\boldsymbol{u})=F(\boldsymbol{u}^{(k-1)})$ 相交于一个椭圆（图中虚线）。由于 $\boldsymbol{u}^{(k-1)}$ 是在通过 $\boldsymbol{u}^{(k-2)}$ 沿 $\boldsymbol{p}^{(k-1)}$ 方向的直线上函数 F 之极小点，所以，$\boldsymbol{p}^{(k-1)}$ 必与该椭圆相切于点 $\boldsymbol{u}^{(k-1)}$。在平面 π_k 上函数 F 取极小值之点显然是这一椭圆的中心 M，所以，我们应取图中通过中心点 M 的方向 $\boldsymbol{p}^{(k)}$ 来作为修正方向。这一方向就是对于该椭圆来说与切线 $\boldsymbol{p}^{(k-1)}$ 共轭的方向，它们将满足如下共轭关系式（见[6] §68）：

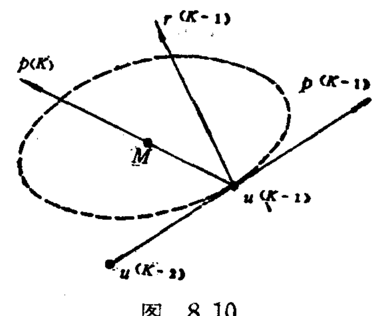


图 8.10

$$(\boldsymbol{A}\boldsymbol{p}^{(k)},\ \boldsymbol{p}^{(k-1)})=(\boldsymbol{p}^{(k)},\ \boldsymbol{A}\boldsymbol{p}^{(k-1)})=0\tag{8.2.77}$$

由于 $\boldsymbol{p}^{(k)}$ 属于平面 π_k，所以，它可表为 $\boldsymbol{r}^{(k-1)}$ 和 $\boldsymbol{p}^{(k-1)}$ 的线性组合，且 $\boldsymbol{r}^{(k-1)}$ 的系数必不等于零。将其适当归一化后 $\boldsymbol{p}^{(k)}$ 必定可以表为如下线性组合形式：

$$\boldsymbol{p}^{(k)}=\boldsymbol{r}^{(k-1)}+e_{k-1}\cdot\boldsymbol{p}^{(k-1)}\quad(k=2,\ 3,\ 4,\ \cdots)\tag{8.2.78}$$

同时，从 $\boldsymbol{p}^{(k)}$ 与 $\boldsymbol{p}^{(k-1)}$ 满足共轭关系(8.2.77)，即可推得：

$$e_{k-1}=-\frac{(\boldsymbol{r}^{(k-1)},\ \boldsymbol{A}\boldsymbol{p}^{(k-1)})}{(\boldsymbol{p}^{(k-1)},\ \boldsymbol{A}\boldsymbol{p}^{(k-1)})}\quad(k=2,\ 3,\ 4,\ \cdots)\tag{8.2.79}$$

在这样确定的修正方向 $\boldsymbol{p}^{(k)}$ 上找出极小点来，就得到如下关系：

$$\begin{aligned}&\boldsymbol{u}^{(k)}=\boldsymbol{u}^{(k-1)}+q_k\boldsymbol{p}^{(k)}\\ &q_k=\frac{(\boldsymbol{r}^{(k-1)},\ \boldsymbol{p}^{(k)})}{(\boldsymbol{A}\boldsymbol{p}^{(k)},\ \boldsymbol{p}^{(k)})}\quad(k=2,\ 3,\ 4,\ \cdots)\end{aligned}\tag{8.2.80}$$

以上这些就是共轭斜量法的主要计算步骤。实际计算中，常将上述公式进行一些简化，从而得到一个形式上更为简单与对称的计算公式。

首先，从(8.2.80)有：

$$\boldsymbol{r}^{(k)}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}^{(k)}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}^{(k-1)}-q_k\boldsymbol{A}\boldsymbol{p}^{(k)}=\boldsymbol{r}^{(k-1)}-q_k\cdot\boldsymbol{A}\boldsymbol{p}^{(k)}\tag{8.2.81}$$

这一公式可以用来递推地计算 $\boldsymbol{r}^{(k)}$。由于 $\boldsymbol{A}\boldsymbol{p}^{(k)}$ 在计算 q_k 时已算得，所以不必将 $\boldsymbol{u}^{(k)}$ 再代入方程中去计算 $\boldsymbol{r}^{(k)}$，因而节省一些计算工作量。

其次，大家知道 $\boldsymbol{u}^{(k)}$ 是函数 F 在整个平面 π_k 内的极小点；向量 $\boldsymbol{r}^{(k)}$ 为函数 F 在点 $\boldsymbol{u}^{(k)}$

处的负梯度向量。从前一事实我们得知平面 π_k 必与椭球面 $F(\boldsymbol{u})=F(\boldsymbol{u}^{(k)})$ 相切于点 $\boldsymbol{u}^{(k)}$。从后一事实得知 $\boldsymbol{r}^{(k)}$ 必与函数 F 的等值曲面 $F(\boldsymbol{u})=F(\boldsymbol{u}^{(k)})$ 正交于点 $\boldsymbol{u}^{(k)}$，即 $\boldsymbol{r}^{(k)}$ 将与平面 π_k 正交。这样便有

$$\begin{aligned}&(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k-1)})=0\\&(\boldsymbol{r}^{(k)}, \boldsymbol{p}^{(k-1)})=0 \quad (k=2, 3, 4, \cdots)\\&(\boldsymbol{r}^{(k)}, \boldsymbol{p}^{(k)})\ \ =0\end{aligned} \tag{8.2.82}$$

利用这些关系，即可将计算 e_{k-1} 及 q_k 的公式进行简化。

由(8.2.82)、(8.2.78)知:

$$(\boldsymbol{r}^{(k-1)}, \boldsymbol{p}^{(k)})=(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})+e_{k-1}(\boldsymbol{r}^{(k-1)}, \boldsymbol{p}^{(k-1)})=(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})$$

所以，从(8.2.80)有:

$$q_k=\frac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{A}\boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)})} \quad (k=1, 2, \cdots) \tag{8.2.83}$$

由(8.2.83)和 $\boldsymbol{A}$ 的正定性可知，当 $\boldsymbol{r}^{(k-1)}\neq 0$ 时，总有 $q_k>0$。这样一来，从(8.2.81)、(8.2.82)又可得:

$$\boldsymbol{A}\boldsymbol{p}^{(k-1)}=\frac{-1}{q_{k-1}}(\boldsymbol{r}^{(k-1)}-\boldsymbol{r}^{(k-2)}) \tag{8.2.84}$$

以及

$$\begin{aligned}(\boldsymbol{r}^{(k-1)}, \boldsymbol{A}\boldsymbol{p}^{(k-1)})&=\frac{-1}{q_{k-1}}[(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})-(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-2)})]\\&=\frac{-1}{q_{k-1}}(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})\\(\boldsymbol{p}^{(k-1)}, \boldsymbol{A}\boldsymbol{p}^{(k-1)})&=\frac{1}{q_{k-1}}(\boldsymbol{p}^{(k-1)}, \boldsymbol{r}^{(k-2)})=\frac{1}{q_{k-1}}(\boldsymbol{r}^{(k-2)}, \boldsymbol{r}^{(k-2)})\end{aligned}$$

所以，我们可将计算 e_{k-1} 的公式(8.2.79)化为:

$$e_{k-1}=\frac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{r}^{(k-2)}, \boldsymbol{r}^{(k-2)})} \quad (k=2, 3, \cdots) \tag{8.2.85}$$

同样，当 $\boldsymbol{r}_{k-1}\neq 0$ 时，总有 $e_{k-1}>0$。

综上所述，利用(8.2.83)、(8.2.80)、(8.2.81)、(8.2.85)和(8.2.78)，我们可将共轭斜量法的计算公式归纳如下:

(1) 任取初始向量 $\boldsymbol{u}^{(0)}$，并计算

$$\boldsymbol{r}^{(0)}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}^{(0)}, \quad \boldsymbol{p}^{(1)}=\boldsymbol{r}^{(0)}$$

(2) 对于 $k=1, 2, \cdots$ 重复如下计算:

$$\begin{cases}q_k=\dfrac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{A}\boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)})}\\ \boldsymbol{u}^{(k)}=\boldsymbol{u}^{(k-1)}+q_k\cdot\boldsymbol{p}^{(k)}\\ \boldsymbol{r}^{(k)}=\boldsymbol{r}^{(k-1)}-q_k\cdot\boldsymbol{A}\boldsymbol{p}^{(k)}\\ e_k=\dfrac{(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)})}{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}\\ \boldsymbol{p}^{(k+1)}=\boldsymbol{r}^{(k)}+e_k\cdot\boldsymbol{p}^{(k)}\end{cases} \quad \text{或者} \quad \begin{cases}\left.\begin{aligned}&e_{k-1}=\dfrac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{r}^{(k-2)}, \boldsymbol{r}^{(k-2)})}\\&\boldsymbol{p}^{(k)}=\boldsymbol{r}^{(k-1)}+e_{k-1}\cdot\boldsymbol{p}^{(k-1)}\end{aligned}\right\}(k\geqslant 2)\\ q_k=\dfrac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{A}\boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)})}\\ \boldsymbol{u}^{(k)}=\boldsymbol{u}^{(k-1)}+q_k\cdot\boldsymbol{p}^{(k)}\\ \boldsymbol{r}^{(k)}=\boldsymbol{r}^{(k-1)}-q_k\cdot\boldsymbol{A}\boldsymbol{p}^{(k)}\end{cases} \tag{8.2.86}$$

(三)共轭斜量法的基本性质

共轭斜量法(8.2.86)的如下基本性质对于理解和使用它是有重要意义的。我们将这些性质归纳为两个定理。

定理 2.14　共轭斜量法(8.2.86)中，逐次修正向量 $\boldsymbol{p}^{(k)}(k=1, 2, \cdots)$，形成一个共轭向量组，即满足关系：

$$(\boldsymbol{p}^{(i)}, \boldsymbol{A}\boldsymbol{p}^{(j)})=0 \quad (i\neq j) \tag{8.2.87}$$

逐次剩余向量 $\boldsymbol{r}^{(k)}(k=0, 1, 2, \cdots)$形成一个正交向量组，即：

$$(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)})=0 \quad (i\neq j) \tag{8.2.88}$$

根据这个定理可以得知，$n+1$ 个向量 $\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \cdots, \boldsymbol{r}^{(n)}$ 中必有一个零向量。因为它们是 n 维空间中的正交向量组，其中非零向量的个数是不可能超过 n 个的。若某个 $\boldsymbol{r}^{(k)}=0$，自然其相应向量 $\boldsymbol{u}^{(k)}$ 就是要求的解向量 $\boldsymbol{u}^*$。因而，又得到下列定理。

定理 2.15　按照共轭斜量法(8.2.86)最多计算 n 步(n 为方程组的阶数)，便可得到方程组(8.2.1)的真解 $\boldsymbol{u}^*$。

定理 2.14 的证明可以用归纳法来进行，下面我们将其简述之。

归纳法的假设为：

在第 k 步$(k\geqslant 1)$计算后，如下关系成立：

$$(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)})=0 \quad i\neq j(0\leqslant i, j\leqslant k) \tag{8.2.89}$$

$$(\boldsymbol{p}^{(i)}, \boldsymbol{A}\boldsymbol{p}^{(j)})=0 \quad i\neq j(1\leqslant i, j\leqslant k) \tag{8.2.90}$$

且 $\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \cdots, \boldsymbol{r}^{(k)}$ 均不为零向量。

归纳法的结论为：

向量 $\boldsymbol{r}^{(k+1)}$、$\boldsymbol{p}^{(k+1)}$ 应满足如下关系：

$$(\boldsymbol{p}^{(k+1)}, \boldsymbol{A}\boldsymbol{p}^{(j)})=0 \quad (j=1, 2, \cdots, k) \tag{8.2.91}$$

$$(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(j)})=0 \quad (j=0, 1, 2, \cdots, k) \tag{8.2.92}$$

证明：当 $j=k$ 时，按照 $\boldsymbol{p}^{(k+1)}$ 的定义(8.2.78)和(8.2.79)容易推知(8.2.91)是成立的。当 $1\leqslant j<k$ 时，从(8.2.86)和(8.2.90)得知：

$$(\boldsymbol{p}^{(k+1)}, \boldsymbol{A}\boldsymbol{p}^{(j)})=(\boldsymbol{r}^{(k)}, \boldsymbol{A}\boldsymbol{p}^{(j)})+e_k(\boldsymbol{p}^{(k)}, \boldsymbol{A}\boldsymbol{p}^{(j)})=(\boldsymbol{r}^{(k)}, \boldsymbol{A}\boldsymbol{p}^{(j)})$$

再从(8.2.84)即得：

$$(\boldsymbol{p}^{(k+1)}, \boldsymbol{A}\boldsymbol{p}^{(j)})=\frac{1}{q_j}[(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(j)})-(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(j-1)})]$$

由归纳法假设和 $q_j>0$，立即可得上式为零，于是(8.2.91)得证。

同时，由共轭斜量法的计算公式(8.2.86)，可以得到：

$$\begin{aligned}(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(j)})&=(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(j)})-q_{k+1}(\boldsymbol{A}\boldsymbol{p}^{(k+1)}, \boldsymbol{r}^{(j)})\\&=(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(j)})-q_{k+1}(\boldsymbol{A}\boldsymbol{p}^{(k+1)}, \boldsymbol{p}^{(j+1)}-e_j\boldsymbol{p}^{(j)})\end{aligned} \tag{8.2.93}$$

当 $j=k$ 时，由此立即推得：

$$(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(k)})=(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)})-q_{k+1}(\boldsymbol{A}\boldsymbol{p}^{(k+1)}, \boldsymbol{p}^{(k+1)})=0$$

当 $1\leqslant j<k$ 时，(8.2.93)变为：

$$\begin{aligned}(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(j)})&=-q_{k+1}[(\boldsymbol{A}\boldsymbol{p}^{(k+1)}, \boldsymbol{p}^{(j+1)})-e_j(\boldsymbol{A}\boldsymbol{p}^{(k+1)}, \boldsymbol{p}^{(j)})]\\&=-q_{k+1}[(\boldsymbol{p}^{(k+1)}, \boldsymbol{A}\boldsymbol{p}^{(j+1)})-e_j(\boldsymbol{p}^{(k+1)}, \boldsymbol{A}\boldsymbol{p}^{(j)})]\end{aligned}$$

由前面已证明的(8.2.91)式，我们立即得知上式为零。

当 $j=0$ 时，因为 $\boldsymbol{p}^{(1)}=\boldsymbol{r}^{(0)}$，所以(8.2.93)变为：

$$(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(0)})=-q_{k+1}(\boldsymbol{A}\boldsymbol{p}^{(k+1)}, \boldsymbol{p}^{(1)})$$

从(8.2.91)亦知其为零,这样,(8.2.92)就得以证明。

此外,当 $k=1$ 时,读者易于直接验证(8.2.89)的正确性,至于(8.2.90)则无需验证。所以定理的结论当 $k=1$ 时是成立的。这样就完成了定理 2.14 的证明。

上述两个定理说明,共轭斜量法本质上是一种直接解法。它在有限步(不多于 n 步)内即可求得真解。实际计算中,由于有舍入误差存在,逐次剩余向量 $\boldsymbol{r}^{(k)}$ 间不能精确满足正交关系,所以,一般来说 $\boldsymbol{r}^{(n)}\neq 0$。并且,系数矩阵愈病态,$\boldsymbol{r}^{(n)}$ 偏离零向量愈远。但是,共轭斜量法的计算公式具有迭代格式的特点,我们也可以把它看作一个迭代解法。如果计算 n 步后 $\boldsymbol{r}^{(n)}\neq 0$,还可用 $\boldsymbol{u}^{(n)}$ 作为向量 $\boldsymbol{u}^{(k-1)}$,继续计算下去。由于每计算一步时只要 $\boldsymbol{r}^{(k)}\neq 0$,函数 F 的值都将减少,所以,这样继续计算下去,直到由于舍入误差影响使得解答不能改进为止,总可以得到一个更好的近似解。此外,许多情况下系数矩阵并不十分病态,我们往往不需要计算 n 步,余量 $\boldsymbol{r}^{(k)}$ 便已充分小,这时就可停止计算,而将对应的向量 $\boldsymbol{u}^{(k)}$ 作为近似解。应该指出,余量 $\boldsymbol{r}^{(k)}$ 的模通常不是单调下降的。尽管如此,实际计算时,一般还是根据余量的大小来判断是否停止计算。

对于系数矩阵为高阶稀疏矩阵的问题,共轭斜量法是一个有效的方法。由于主要工作量是需要完成 $\boldsymbol{A}\cdot\boldsymbol{z}$ 型的运算,所以,不必存放系数矩阵 $\boldsymbol{A}$ 的全部元素,仅需保存或由机器在计算过程中临时产生其非零元素即可。同时,计算经验表明,对于不是非常病态的问题,此法收敛较快。一般来说,所需迭代的次数远小于矩阵的阶数 n。对于比较病态的问题,重复足够多次迭代(迭代次数有时可能等于阶数 n 的 3~5 倍)后,一般也能得到满意结果。特别是使用共轭斜量法时,不需要估计矩阵特征值的上下界和某些迭代参数,这也是一个重要的优点。其主要缺点是需要 $3n$(或 $4n$)个工作单元来存放逐次的向量 $\boldsymbol{r}^{(k)}$、$\boldsymbol{p}^{(k)}$、$\boldsymbol{u}^{(k)}$(或$\boldsymbol{A}\boldsymbol{p}^{(k)}$)。此外,每步的计算量也较大些。后一缺点往往由其迭代次数的减少而得到弥补。在目前计算机存储量已不是很小的情况下,前一缺点也不是十分重要的,特别是比起其它直接法来说更是如此。目前,共轭斜量法是高阶稀疏矩阵问题中的比较常用的方法之一。

共轭斜量法的算法语言程序见本章最后所附程序 13。

§8.3 线性矛盾方程组的最小二乘解法

本节简单讨论一下在计算机上行之有效的解线性矛盾方程组的方法。这类问题在实践中是经常遇到的。例如,从 n 个已知值 d_i(测量值或其它)决定 m 个未知量 x_j 的问题,只要已知值的个数 n 大于唯一决定未知量 x_j 所必需的方程数目 m,就会得到一个矛盾方程组。在一定条件下,把所得的方程组线性化,就得到线性矛盾方程组。大地测量问题、曲线拟合问题以及某些统计计算问题等,均属于这个类型。

未知量 x_j 与已知值 d_i 本来应该严格满足某些方程式,但由于已知值总是包含有误差,并且方程个数多于未知数个数,所以这些方程式中必有"矛盾"。为了解决"矛盾",在确定 x_j 时需要考虑各方程的剩余量,并使这些剩余量符合某种意义下的极小性。一般可以采取使剩余量的最大绝对值为最小的原则(即切比雪夫原则)或使剩余量的平方和为最小的原则(即高斯原则,又称为最小二乘原则)来决定未知数 x_j。由于已知值中的误差(例如测量误差等)通常将满足所谓高斯分布,此时,用最小二乘原则求得的解答 x_j 将具有最小方差。所以,从概率论的角度来说,最小二乘原则是有其根据的。同时,从计算的角度来看,最小二乘原则的处理也更为简单和易于接受,故我们这里只讨论最小二乘原则。

8.3.1 法方程组的建立

假定我们要求解如下线性矛盾方程组:

$$\begin{cases} c_{11}x_1+c_{12}x_2+\cdots+c_{1m}x_m+d_1=0 \\ c_{21}x_1+c_{22}x_2+\cdots+c_{2m}x_m+d_2=0 \\ \cdots\cdots\cdots\cdots \\ c_{n1}x_1+c_{n2}x_2+\cdots+c_{nm}x_m+d_n=0 \end{cases} \quad (n>m) \tag{8.3.1}$$

其中 d_i 为已知值,系数 c_{ij} 亦为已知量, x_j 为待定的未知数,采用矩阵符号可以表为:

$$\boldsymbol{c}\boldsymbol{x}+\boldsymbol{d}=0 \tag{8.3.2}$$

矩阵 $\boldsymbol{c}$ 为 $n\times m$ 长方阵, 其元素为 c_{ij}, 并且假定矩阵 $\boldsymbol{c}$ 的秩为 m。$\boldsymbol{x}$ 为待求的 m 维解向量 $(x_1, x_2, \cdots, x_m)^T$; $\boldsymbol{d}$ 为已知的 n 维向量$(d_1, d_2, \cdots, d_n)^T$。

由于上述方程组中包含着"矛盾", 因而, 严格说来对于任何 $\boldsymbol{x}$(8.3.1)均不会成为真正的等式。为此,我们引入剩余量,并把方程写为:

$$\boldsymbol{c}\boldsymbol{x}+\boldsymbol{d}=\boldsymbol{f}, \quad \boldsymbol{f}=(f_1, f_2, \cdots, f_n)^T \tag{8.3.3}$$

一般来说,有无穷多种 $\boldsymbol{f}$ 的取法均可找到相应的 $\boldsymbol{x}$,使上式成为严格的等式。所谓采用最小二乘原则来决定向量 $\boldsymbol{x}$, 就是要求找出使 $\boldsymbol{f}^T\cdot\boldsymbol{f}$ 最小的解答 $\boldsymbol{x}$。 这时, $\boldsymbol{x}$ 将唯一地确定, 因而相应的 $\boldsymbol{f}$ 亦唯一确定。 我们就把这样求得的解向量 $\boldsymbol{x}$ 称为(8.3.1)的最小二乘解。 从 $\boldsymbol{f}^T\cdot\boldsymbol{f}=\min$ 的条件可以导出最小二乘解 $\boldsymbol{x}$ 所应满足的方程组。通常称这种方程组为法方程组。

现在就来建立法方程组,这时有:

$$\sum_{i=1}^{n} f_i^2=\boldsymbol{f}^T\cdot\boldsymbol{f}=(\boldsymbol{c}\boldsymbol{x}+\boldsymbol{d})^T\cdot(\boldsymbol{c}\boldsymbol{x}+\boldsymbol{d})=\boldsymbol{x}^T\boldsymbol{c}^T\cdot\boldsymbol{c}\boldsymbol{x}+2\boldsymbol{x}^T\boldsymbol{c}^T\boldsymbol{d}+\boldsymbol{d}^T\boldsymbol{d} \tag{8.3.4}$$

这是一个 x_j 的二次函数, 如果我们求出这个函数的各个偏微商, 令其为零, 并仿照 §8.2.5 的讨论,便可得到要求的法方程组,但是这里有更简单的处理办法。靠直接计算便容易验证(8.3.4)右端应等于:

$$(\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{x}+\boldsymbol{c}^T\boldsymbol{d})^T\cdot(\boldsymbol{c}^T\boldsymbol{c})^{-1}\cdot(\boldsymbol{c}^T\boldsymbol{x}+\boldsymbol{c}^T\boldsymbol{d})-\boldsymbol{d}^T\boldsymbol{c}(\boldsymbol{c}^T\boldsymbol{c})^{-1}\cdot\boldsymbol{c}^T\boldsymbol{d}+\boldsymbol{d}^T\boldsymbol{d}$$

同时, 可以证明当矩阵 $\boldsymbol{c}$ 之秩为 m 时, $\boldsymbol{c}^T\boldsymbol{c}$ 为正定矩阵。 这是由于$(\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{x}, \boldsymbol{x})=(\boldsymbol{c}\boldsymbol{x}, \boldsymbol{c}\boldsymbol{x})$为非负二次型,并且仅当 $\boldsymbol{c}\boldsymbol{x}=0$ 时才为零。但因 $\boldsymbol{c}$ 之秩为 m, 从 $\boldsymbol{c}\boldsymbol{x}=0$ 将得出 $\boldsymbol{x}=0$。所以可以得知$(\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{x}, \boldsymbol{x})$为正定二次型, 即 $\boldsymbol{c}^T\boldsymbol{c}$ 因而$(\boldsymbol{c}^T\boldsymbol{c})^{-1}$为正定矩阵。 这样一来, 由上式的后两项与 $\boldsymbol{x}$ 无关,第一项又应非负, 知其极小值必在其第一项为零, 即 $\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{x}+\boldsymbol{c}^T\boldsymbol{d}=0$ 时达到。于是我们便得到使 $\boldsymbol{f}^T\cdot\boldsymbol{f}$ 达极小的未知量 $\boldsymbol{x}$ 亦即最小二乘解 $\boldsymbol{x}$ 应满足的法方程组:

$$\boldsymbol{c}^T\boldsymbol{c}\boldsymbol{x}+\boldsymbol{c}^T\boldsymbol{d}=0 \tag{8.3.5}$$

8.3.2 法方程组的求解

法方程组的系数矩阵 $\boldsymbol{A}=\boldsymbol{c}^T\cdot\boldsymbol{c}$ 是对称正定的,因而,本章 §8.1 与 §8.2 中所讨论的很多方法都可以用来求解它。例如, §8.1 讨论的各种直接法中,我们可以采用平方根法, §8.2 讨论的各种迭代法中,可以采用逐次松弛法和共轭斜量法等。但是,由于法方程组本身的某些特点,直接使用上述方法有时会得出不好的结果。因此,还需对如何求解法方程组的问题作进一步的讨论。

法方程组(8.3.5)区别于一般方程组的特点有两个。其一是形成其系数矩阵 $\boldsymbol{A}$ 时需作矩阵乘积 $\boldsymbol{c}^T\cdot\boldsymbol{c}$, 这不仅会引入某些舍入误差,增加一些运算量,而且有时还会破坏原来矩阵 $\boldsymbol{c}$ 的某些有用的特殊形状。其二是其系数矩阵 $\boldsymbol{A}=\boldsymbol{c}^T\cdot\boldsymbol{c}$ 往往是对于解方程组来说的病态矩阵。前一个特点是明显的,后一特点可从下述例子与定性说明中得知。

〔例子〕 考虑从 $N+1$ 个给定点上的函数值 φ_i 去确定一个 $n(<N)$ 阶多项式的问题。如果给定的点为$(N=11)$:

$$x_i=-6,\ -5,\ \cdots,\ -1,\ 1,\ 2,\ \cdots,\ 5,\ 6$$

要求确定 8 次多项式 $\varphi(x)=a_0+a_1x+a_2x^2+\cdots+a_8x^8$ 的系数 a_k, 则应有:

$$a_0+a_1x_i+a_2x_i^2+\cdots+a_8x_i^8=\varphi_i \quad (i=1,\ 2,\ \cdots,\ 12)$$

或者写为:

$$\boldsymbol{ca}=\boldsymbol{\varphi}$$

相应法方程为:

$$\boldsymbol{Aa}=\boldsymbol{c}^T\boldsymbol{ca}=\boldsymbol{c}^T\boldsymbol{\varphi}$$

其中矩阵 $\boldsymbol{c}$ 之元素 $c_{ij}=x_i^{j-1}$。

这样便有如下条件数估计:

$$P(\boldsymbol{A})=P(\boldsymbol{c}^T\boldsymbol{c})=\frac{\lambda_{\max}(\boldsymbol{c}^T\boldsymbol{c})}{\lambda_{\min}(\boldsymbol{c}^T\boldsymbol{c})}\geqslant\frac{\max\limits_k(a_{kk})}{\min\limits_k(a_{kk})}\sim 4.96\times10^{11}$$

所以,按照§8.1 所列举的误差估计式(8.1.59), 在字长为十一位十进制数的计算机上求解上述问题的法方程组 $\boldsymbol{c}^T\boldsymbol{ca}=\boldsymbol{c}^T\boldsymbol{\varphi}$ 时,可能会得出完全错误的结果。

显然,上述例子中的法方程组可以说是"病态的",因而, 其求解有一定困难。原来的矛盾方程组并不十分"病态",为什么其法方程组变为"病态的"呢? 这里仅给予一些粗略说明。

原来的矛盾方程组 $\boldsymbol{cx}+\boldsymbol{d}=0$ 的病态程度(即将其系数矩阵或自由项作微小改变后,其解答变化的程度)在一定条件下可用下列方式定义的条件数 $\varkappa(\boldsymbol{c})$ 来近似地衡量:

$$\varkappa(c)=\max_{\|x\|_2=1}\|cx\|_2/\min_{\|x\|_2=1}\|cx\|_2=\mu_1/\mu_n \tag{8.3.6}$$

其中 $\mu_1=\sqrt{\lambda\max(c^Tc)}$, $\mu_n=\sqrt{\lambda\min(c^Tc)}$。

也就是说,若给矩阵 $\boldsymbol{c}$ 及自由项 $\boldsymbol{d}$ 以微小变化 $\delta\boldsymbol{c}$ 和 $\delta\boldsymbol{d}$, 则在一定条件下,其最小二乘解 $\boldsymbol{x}$ 的变化 $\delta\boldsymbol{x}$ 将满足下列形式的不等式:

$$\|\delta\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2\leqslant\varkappa(\boldsymbol{c})\left(k_1\frac{\|\delta\boldsymbol{c}\|_2}{\|\boldsymbol{c}\|_2}+k_2\frac{\|\delta\boldsymbol{d}\|_2}{\|\boldsymbol{d}\|_2}\right) \tag{8.3.7}$$

但是,若将问题化为求解法方程组时,可以验证其系数矩阵 $\boldsymbol{A}$ 的条件数将为:

$$P(\boldsymbol{A})=P(\boldsymbol{c}^T\boldsymbol{c})=(\mu_1/\mu_n)^2=\varkappa(\boldsymbol{c})^2 \tag{8.3.8}$$

所以,一般来说, 化为法方程组后其系数矩阵的条件数将变为原来的平方, 即"病态"程度将大大增加。正是由于这个原因, 许多本来不太"病态"的问题, 化为法方程后就变得相当病态。自然, 上述的说明和所得结论均是很粗略的。详细的严格论证, 可以参看[22]的第五段。

由于法方程组的这一特点, 有必要采用具有更高数值稳定性的计算方法来解决问题。目前,一般使用镜像映射法和正交化法, 并已获得了很好的效果。下面将讨论这两个方法。此外,实践中也有许多问题的法方程组并不是十分病态的,这时,仍可使用§8.1 或§8.2 中的某些解法。为了避免得出 $\boldsymbol{c}^T\cdot\boldsymbol{c}$ 的运算,此时我们可将原来的计算公式加以变形。共轭斜

量法最适宜于达到这个目的，计算实践也证明其数值效果良好。它将在正交化法之后给以讨论。至于其它解法，可以参见本章的前两节。

(一)镜像映射法

我们知道，正交变换下任何向量的欧氏长度是不改变的。因而，若 $\boldsymbol{Q}$ 为 $n\times n$ 正交矩阵，则有：

$$\|\boldsymbol{d}+\boldsymbol{c}\boldsymbol{x}\|_2=\|\boldsymbol{Q}\cdot(\boldsymbol{d}+\boldsymbol{c}\boldsymbol{x})\|_2=\|\boldsymbol{Q}\boldsymbol{d}+\boldsymbol{Q}\boldsymbol{c}\boldsymbol{x}\|_2 \tag{8.3.9}$$

于是，寻求使 $\|\boldsymbol{d}+\boldsymbol{c}\boldsymbol{x}\|_2$ 极小的解答 $\boldsymbol{x}$ 的问题与求出使 $\|\boldsymbol{Q}\boldsymbol{d}+\boldsymbol{Q}\boldsymbol{c}\boldsymbol{x}\|_2$ 极小的解答是等价的，而这一解答正是我们所要求的最小二乘解 $\boldsymbol{x}$。

如果我们能够找到一个正交矩阵 $\boldsymbol{Q}$，使得：

$$\boldsymbol{Q}\cdot\boldsymbol{c}\widetilde{\boldsymbol{R}}==\begin{pmatrix}\boldsymbol{R}\\ \cdots\\ 0\end{pmatrix}^{m\times m}_{(n-m)\times m} \tag{8.3.10}$$

其中 $\boldsymbol{R}$ 为 $m\times m$ 上三角型矩阵。再将 $\tilde{\boldsymbol{d}}=\boldsymbol{Q}\cdot\boldsymbol{d}$ 的前 m 个分量记为 $\boldsymbol{e}$，后 $(n-m)$ 个分量记为 $\boldsymbol{g}$，那么便有：

$$\|\boldsymbol{d}+\boldsymbol{c}\boldsymbol{x}\|_2=\left\|\begin{pmatrix}\boldsymbol{R}\\0\end{pmatrix}\cdot\boldsymbol{x}+\begin{pmatrix}\boldsymbol{e}\\\boldsymbol{g}\end{pmatrix}\right\|_2=[(\boldsymbol{R}\boldsymbol{x}+\boldsymbol{e})^T\cdot(\boldsymbol{R}\boldsymbol{x}+\boldsymbol{e})+\boldsymbol{g}^T\cdot\boldsymbol{g}]^{1/2} \tag{8.3.11}$$

显然，(8.3.11)的极小值应在 $\boldsymbol{x}=-\boldsymbol{R}^{-1}\cdot\boldsymbol{e}$ 时达到，且极小值为：$(\boldsymbol{g}^T\cdot\boldsymbol{g})^{1/2}$。由于 $\boldsymbol{R}$ 为 $m\times m$ 上三角型矩阵，所以，计算 $\boldsymbol{R}^{-1}\cdot\boldsymbol{e}$ 是很简单的[见(8.1.7)]。于是，不需经过法方程组(8.3.5)，而只要直接解一次三角型方程组 $\boldsymbol{R}\boldsymbol{x}+\boldsymbol{e}=0$，便可求出所要求的解答 $\boldsymbol{x}$。

满足(8.3.10)要求的正交矩阵 $\boldsymbol{Q}$ 比较容易找到。§8.1 中讨论的镜像映射矩阵即可用来达到此目的。以上这些就是镜像映射法的基本概念。按照 8.1.6 节中的具体作法，我们找出一系列镜像映射矩阵 $\boldsymbol{H}_i$ 来左乘矩阵 $\boldsymbol{c}$，即可将其化为(8.3.10)的形状。只是现在要处理的矩阵 $\boldsymbol{c}$ 为 $n\times m$ 的长方阵，因而逐次变换矩阵 $\boldsymbol{H}_i$ 仍为 $n\times n$ 矩阵，但变换总次数则为 m 次。所以，经 m 次变换后我们就有：

$$\begin{aligned}\widetilde{\boldsymbol{R}}&=\boldsymbol{H}_m\cdot\boldsymbol{H}_{m-1}\cdot\cdots\cdot\boldsymbol{H}_2\cdot\boldsymbol{H}_1\cdot\boldsymbol{c}=\begin{bmatrix}\boldsymbol{R}\\\cdots\\0\end{bmatrix}^{m\times m}_{(n-m)\times n}\\ \tilde{\boldsymbol{d}}&=\boldsymbol{H}_m\cdot\boldsymbol{H}_{m-1}\cdot\cdots\cdot\boldsymbol{H}_2\cdot\boldsymbol{H}_1\cdot\boldsymbol{d}=\begin{bmatrix}\boldsymbol{e}\\\cdots\\\boldsymbol{g}\end{bmatrix}^{m}_{n-m}\end{aligned} \tag{8.3.12}$$

这样，$\boldsymbol{Q}=\boldsymbol{H}_m\cdot\boldsymbol{H}_{m-1}\cdot\cdots\cdot\boldsymbol{H}_2\cdot\boldsymbol{H}_1$。若令 $\boldsymbol{c}=\boldsymbol{c}^{(1)}$、$\boldsymbol{d}=\boldsymbol{d}^{(1)}$，完全仿照(8.1.34)式的符号记法和推导过程，即可得出如下计算公式：

$$\left\{\begin{aligned}&\alpha_k=\left[\sum_{l=k}^{n}(c_{lk}^{(k)})^2\right]^{1/2}\\&\boldsymbol{u}_k=(0,\ \cdots,\ 0,\ c_{kk}^{(k)}+\operatorname{sign}(c_{kk}^{(k)})\cdot\alpha_k,\ c_{k+1,k}^{(k)},\cdots,\ c_{nk}^{(k)})^T\\&\sigma_k=2\alpha_k\cdot(\alpha_k+|c_{kk}^{(k)}|)\\&\boldsymbol{q}_k^T=2\boldsymbol{u}_k^T\cdot\boldsymbol{c}^{(k)}/\sigma_k\\&\mu_k=2\boldsymbol{u}_k^T\boldsymbol{d}^{(k)}/\sigma_k\\&\boldsymbol{c}^{(k+1)}=\boldsymbol{c}^{(k)}-\boldsymbol{u}_k\cdot\boldsymbol{q}_k^T\\&\boldsymbol{d}^{(k+1)}=\boldsymbol{d}^{(k)}-\mu_k\boldsymbol{u}_k\end{aligned}\right. \tag{8.3.13}$$

$$(k=1,\ 2,\ \cdots,\ m)$$

最后即有: $$\widetilde{R}=c^{(m+1)},\ \tilde{d}=d^{(m+1)}$$

注意到 $\tilde{d}$ 的后 $n-m$ 个分量的平方和就是(8.3.3)中的 $\sum_{i=1}^{n} f_i^2$ 的极小值[参见(8.3.11)],在求解过程中我们很容易同时算出它,无需将解答代入原方程(8.3.3)去重新计算。

本章最后所附的程序 14 便是用镜像映射法解线性矛盾方程组的算法语言程序。

(二) 正交化法

可以用所谓格拉姆-施密特(Gramm-Schmidt)正交化过程(以后简记为 G-S 过程)来求线性矛盾方程组(8.3.1)的最小二乘解 x。

将 G-S 正交化过程用于矩阵 c 的各列 c_i,便可得出一个列正交(且规一化)的同维矩阵 s。其具体步骤如下:

(1) 将 c_1 规一化,作为 s 的第一列 s_1:

$$s_1=c_1/r_{11},\quad r_{11}=\sqrt{c_1^T\cdot c_1}$$

(2) 对于 $k=2, 3, \cdots, n$, 按下列公式求出 s 的第 k 列 s_k:

$$b_k=c_k-\sum_{j=1}^{k-1} r_{jk}s_j;\quad r_{jk}=s_j^T\cdot c_k\quad (j=1,\ 2,\ \cdots,\ k-1)$$

$$s_k=b_k/r_{kk};\quad r_{kk}=\sqrt{b_k^T\cdot b_k}$$

注意到向量 b_k 是与 $s_1, s_2, \cdots, s_{k-1}$ 诸向量正交的。所以, s_k 将是与 $s_1, s_2, \cdots, s_{k-1}$ 正交的规一化向量。

把上述过程改写一下即得:

$$\begin{cases} c_1=r_{11}s_1 \\ c_2=r_{12}s_1+r_{22}s_2 \\ c_3=r_{13}s_1+r_{23}s_2+r_{33}s_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ c_m=r_{1m}s_1+r_{2m}s_2+\cdots+r_{mm}s_m \end{cases}$$

用矩阵符号即可记为:

$$c=S\cdot R \tag{8.3.14}$$

其中, $R=[r_{ij}]$ 为 $m\times m$ 上三角矩阵; S 为 $n\times m$ 列正交(规一)矩阵,即 $S^T\cdot S=I$。

从(8.3.14)可知, G-S 过程用于矩阵 c 之各列时, 相当于将 c 分解为列正交阵和上三角阵的乘积。将分解式(8.3.14)代入法方程组(8.3.5),便得到:

$$R^TS^TSRx+R^T\cdot S^Td=0$$

由于 c 的秩为 m,矩阵 R 应为非奇的,用 $(R^T)^{-1}$ 左乘上式两端即得:

$$Rx+S^T\cdot d=0 \tag{8.3.15}$$

所以,求得 R 和 $S^T\cdot d$ 后,只需要解一次三角形方程组(8.3.15)即可求得最小二乘解 x。

实际使用 G-S 正交化过程时, 发现前述的计算格式(即每次将 c 的一列 c_i 加入到前面求得的正交向量组 s_j 中,并保持未加入的 c_i 各列不变者)数值稳定性较差。因而,目前采用精度高得多的所谓"修改的 G-S 正交化过程"。其具体计算步骤如下:

(1) 将 $c^{(1)}=c$ 的第一列 $c_1^{(1)}$ 规一化作为 S 的第一列 s_1:

$$s_1=c_1^{(1)}/r_{11}\quad r_{11}=\sqrt{(c_1^{(1)},\ c_1^{(1)})}$$

然后,将 $c^{(1)}$ 的第 2 至 m 列分别与 s_1 正交化,得出向量组: $c_2^{(2)}, c_3^{(2)}, \cdots, c_m^{(2)}$。即是说,令:

$$\boldsymbol{c}_j^{(2)}=\boldsymbol{c}_j^{(1)}-r_{1j}\boldsymbol{s}_1;\quad r_{1j}=(\boldsymbol{s}_1,\ \boldsymbol{c}_j^{(1)})\quad (j=2,\ 3,\ \cdots,\ m) \tag{8.3.16}$$

(2) 类似于步骤 (1)，对于 $k=2,\ 3,\ \cdots,\ m$，执行下列运算:

$$\boldsymbol{s}_k=\boldsymbol{c}_k^{(k)}/r_{kk};\quad r_{kk}=\sqrt{(\boldsymbol{c}_k^{(k)},\ \boldsymbol{c}_k^{(k)})}$$

$$\boldsymbol{c}_j^{(k+1)}=\boldsymbol{c}_j^{(k)}-r_{kj}\cdot\boldsymbol{s}_k;\quad r_{kj}=(\boldsymbol{s}_k,\ \boldsymbol{c}_j^{(k)})\quad (j=k+1,\ k+2,\ \cdots,\ m) \tag{8.3.17}$$

便得到正交规一化的向量组 $\boldsymbol{s}_k(k=1,\ 2,\ \cdots,\ m)$。

(3) 计算 $\boldsymbol{S}^T\cdot\boldsymbol{d}$，并解三角形方程组:

$$\boldsymbol{Rx}+\boldsymbol{S}^T\cdot\boldsymbol{d}=0;\quad \boldsymbol{R}=[r_{ij}] \tag{8.3.18}$$

即得最小二乘解 $\boldsymbol{x}=-\boldsymbol{R}^{-1}\cdot\boldsymbol{S}^T\cdot\boldsymbol{d}$。

在计算 $\boldsymbol{s}_k$ 时，也可以在 $\boldsymbol{c}_j^{(k)}(j=k,\ k+1,\ \cdots,\ m)$ 中挑选一个模较大者来作为 $\boldsymbol{c}_k^{(k)}$(进行列交换)。这样，处于分母上的 r_{kk} 将有较大的模，有利于控制舍入误差的增长。这一措施叫做选主列。对于镜像映射法也可作类似处理。计算实践表明，这样作对精确度提高并无太大好处，反而使程序复杂化和增加运算量，所以，实际计算中常常不采用选主列措施。

修改正交化法解线性矛盾方程组的算法语言程序可见本章最后的程序 15。

正交化法或镜像映射法与对于法方程组使用平方根法求解有密切的关系。在表 8.1 中，对照地列举了两个方法的计算步骤。

表 8.1

法方程组平方根法	正交化法
$\boldsymbol{A}=\boldsymbol{c}^T\cdot\boldsymbol{c}$; $\boldsymbol{b}=\boldsymbol{c}^T\cdot\boldsymbol{d}$	
$\boldsymbol{Ax}+\boldsymbol{b}=0$(法方程)	
$\boldsymbol{A}=\boldsymbol{R}^T\cdot\boldsymbol{R}$(平方根分解)	$\boldsymbol{c}=\boldsymbol{S}\cdot\boldsymbol{R}$ } 正交化过程
$\boldsymbol{R}^T\boldsymbol{y}+\boldsymbol{b}=0$(解中间变量)	$\boldsymbol{g}=-\boldsymbol{S}^T\cdot\boldsymbol{d}$ } 正交化过程
$\boldsymbol{Rx}-\boldsymbol{y}=0$(回代)	$\boldsymbol{Rx}-\boldsymbol{g}=0$(回代)

从表中可以看出，两个方法的计算步骤是有类似之处的。实际上可以证明两个方法中的矩阵 $\boldsymbol{R}$ 是相同的，回代过程亦完全一样。证明很简单，若在正交化法中有 $\boldsymbol{c}=\boldsymbol{S}\cdot\boldsymbol{R}_1$，则 $\boldsymbol{A}=\boldsymbol{c}^T\cdot\boldsymbol{c}=\boldsymbol{R}_1^T\cdot\boldsymbol{S}^T\cdot\boldsymbol{S}\cdot\boldsymbol{R}_1=\boldsymbol{R}_1^T\cdot\boldsymbol{R}_1$。但在平方根法中亦有分解式 $\boldsymbol{A}=\boldsymbol{R}_2^T\cdot\boldsymbol{R}_2$。从分解式的唯一性很容易推知: $\boldsymbol{R}_1=\boldsymbol{R}_2$。此外，平方根法中的向量 $\boldsymbol{y}=-(\boldsymbol{R}^T)^{-1}\cdot\boldsymbol{b}=-\boldsymbol{R}^{-T}\cdot\boldsymbol{c}^T\cdot\boldsymbol{d}=-\boldsymbol{R}^{-T}\cdot\boldsymbol{R}^T\cdot\boldsymbol{S}^T\boldsymbol{d}=-\boldsymbol{S}^T\boldsymbol{d}=\boldsymbol{g}$。所以，两者的回代过程也完全相同。这样，我们可以认为两个方法在数学上是等价的。但是，从数值计算的角度来看，两个方法却有本质的差别。前面已经提到，化为法方程组后，系数矩阵的条件数 $P(\boldsymbol{A})=\varkappa(\boldsymbol{c})^2$，故其病态程度大为增加。而在正交化法(或镜像映射法)中，只需要求解一次三角形方程组: $\boldsymbol{Rx}-\boldsymbol{g}=0$，其系数矩阵的条件数 $P(\boldsymbol{R})=[\lambda_{\max}(\boldsymbol{R}^T\cdot\boldsymbol{R})/\lambda_{\min}(\boldsymbol{R}^T\boldsymbol{R})]^{1/2}=[\lambda_{\max}(\boldsymbol{R}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{R})/\lambda_{\min}(\boldsymbol{R}^T\cdot\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{R})]^{1/2}=\varkappa(\boldsymbol{SR})=\varkappa(\boldsymbol{c})$。即保持条件数不变。这样，我们就避免了处理更为病态的法方程组的步骤。所以，正交化方法(或镜像映射法)从数值计算方面来说较为优越一些(注意! 严格说来，前述结论只在与最小二乘解 $\boldsymbol{x}$ 相应的剩余量 $\boldsymbol{f}$ 的模很小时才成立。当原来方程组中的“矛盾”较大，即 $\boldsymbol{f}$ 较大时，正交化法或镜像映射法中所得解 $\boldsymbol{x}$ 的相对误差仍与 $\varkappa(\boldsymbol{c})^2$ 有关，详见[22])。计算实践也表明正交化法和镜像映射法的数值稳定性较高，特别是当剩余量 $\boldsymbol{f}$ 的模较小时(即原方程组中的“矛盾”不大时)正交化法或镜像映射法的精确度比直接解法方程组

的各种直接法高得多(详见[23])。综上所述,在解线性矛盾方程组时,应该十分注意问题的提出和数据的加工,以便使得所形成的矛盾方程组中"矛盾"尽量地小。然后,再用镜像映射法或正交化法求解所得的矛盾方程组。这样,一般来说是可以获得满意结果的。

最后还要指出,利用正交化法中的分解式 $\boldsymbol{c}=\boldsymbol{SR}$,还可以求出一些有用的结果。例如:

$$
\begin{aligned}
&\det(\boldsymbol{c}^T\cdot\boldsymbol{c})=(r_{11}\cdot r_{22}\cdot\cdots\cdot r_{nn})^2\\
&(\boldsymbol{c}^T\cdot\boldsymbol{c})^{-1}=(\boldsymbol{R}^T\boldsymbol{S}^T\cdot\boldsymbol{SR})^{-1}=\boldsymbol{R}^{-1}\cdot\boldsymbol{R}^{-T}
\end{aligned}
\tag{8.3.19}
$$

由于 $\boldsymbol{R}$ 是上三角形矩阵,如下格式还可以同时算出 $\boldsymbol{R}^{-1}$ 及$(\boldsymbol{c}^T\cdot\boldsymbol{c})^{-1}$的元素。为简单起见,现以三阶矩阵为例来加以说明:

令
$$(\boldsymbol{c}^T\cdot\boldsymbol{c})^{-1}=\boldsymbol{X}$$

则从(8.3.19)第二式有:

$$\boldsymbol{RX}=\boldsymbol{R}^{-T}$$

注意到 $\boldsymbol{X}$ 为对称矩阵,$\boldsymbol{R}^{-T}$ 为下三角矩阵,其对角线元为 $1/r_{ii}$,故对于三阶情形,我们有:

$$
\begin{bmatrix} r_{11} & r_{12} & r_{13}\\ 0 & r_{22} & r_{23}\\ 0 & 0 & r_{33}\end{bmatrix}\cdot\begin{bmatrix} x_{11} & x_{12} & x_{13}\\ x_{12} & x_{22} & x_{23}\\ x_{13} & x_{23} & x_{33}\end{bmatrix}=\begin{bmatrix} \frac{1}{r_{11}} & 0 & 0\\ R_{21} & \frac{1}{r_{22}} & 0\\ R_{31} & R_{32} & \frac{1}{r_{33}}\end{bmatrix}
$$

首先可以从下式解出 $\boldsymbol{X}$ 的第3列:

$$
\boldsymbol{R}\cdot\begin{bmatrix} x_{13}\\ x_{23}\\ x_{33}\end{bmatrix}=\begin{bmatrix} 0\\ 0\\ \frac{1}{r_{33}}\end{bmatrix},
$$

然后,应从下式解出 $\boldsymbol{X}$ 的第2列:

$$
\boldsymbol{R}\cdot\begin{bmatrix} x_{12}\\ x_{22}\\ x_{23}\end{bmatrix}=\begin{bmatrix} 0\\ \frac{1}{r_{22}}\\ R_{32}\end{bmatrix}
$$

但由于 x_{23} 已求得(所以 $R_{32}=x_{23}\cdot r_{33}$ 亦为已知),这样,从上式可以直接求出 x_{12}, x_{22}。完全类似地,由于 x_{12}、x_{13} 已求得,故有 $R_{31}=r_{33}\cdot x_{13}$; $R_{21}=r_{22}x_{12}+r_{23}x_{13}$,于是我们可以从下式解出 $\boldsymbol{X}$ 的第一列第一个元素 x_{11}:

$$
\boldsymbol{R}\cdot\begin{bmatrix} x_{11}\\ x_{12}\\ x_{13}\end{bmatrix}=\begin{bmatrix} \frac{1}{r_{11}}\\ R_{21}\\ R_{31}\end{bmatrix}
$$

上述计算 $\det(\boldsymbol{c}^T\boldsymbol{c})$ 和 $(\boldsymbol{c}^T\boldsymbol{c})^{-1}$ 的办法,在某些统计计算问题中是有用处的。

(三)共轭斜量法

8.2.5节中讨论的共轭斜量法可以用来求解法方程组(8.3.5)。当法方程组的系数矩阵不是非常病态时,其收敛速度和精确度均是较高的。特别是当矩阵 $\boldsymbol{c}$ 是高阶的稀疏矩阵时,采取适当修改措施后,共轭斜量法的计算公式中可以避免形成法方程组系数矩阵 $\boldsymbol{c}^T\boldsymbol{c}$ 的运算,而仅需作矩阵 $\boldsymbol{c}$ 与某些向量的乘积。即是说,我们可以保持系数矩阵 $\boldsymbol{c}$ 的稀疏特点,

更有效地处理高阶问题。因而,共轭斜量法在求解线性矛盾方程时是经常使用的。下面,我们就来导出修改后的共轭斜量法的计算公式,并对其作简单讨论。

显然,共轭斜量法中的主要计算量花费在计算 $\boldsymbol{Ap}^{(k)}$ 上,而 $\boldsymbol{Ap}^{(k)}$ 在计算公式中共出现两次(见(8.2.86))。在计算$(\boldsymbol{Ap}^{(k)}, \boldsymbol{p}^{(k)})$时,因 $\boldsymbol{A}=\boldsymbol{c}^T\cdot\boldsymbol{c}$,自然可以化为$(\boldsymbol{cp}^{(k)}, \boldsymbol{cp}^{(k)})$的形式,从而避免算出 $\boldsymbol{c}^T\cdot\boldsymbol{c}$。计算 $\boldsymbol{r}^{(k)}$ 的公式则可换为⊖:

$$\boldsymbol{r}^{(k)}=\boldsymbol{c}^T\cdot\boldsymbol{f}^{(k)}=\boldsymbol{c}^T(\boldsymbol{cx}^{(k)}+\boldsymbol{d}) \tag{8.3.20}$$

由于 $\boldsymbol{x}^{(k)}=\boldsymbol{x}^{(k-1)}+q_k\cdot\boldsymbol{p}^{(k)}$,所以 $\boldsymbol{cx}^{(k)}=\boldsymbol{cx}^{(k-1)}+q_k\cdot\boldsymbol{cp}^{(k)}$。将此式两端均加上 $\boldsymbol{d}$,即可看出 $\boldsymbol{f}^{(k)}$ 就可用如下递推公式计算:

$$\boldsymbol{f}^{(k)}=\boldsymbol{f}^{(k-1)}+q_k(\boldsymbol{c}\cdot\boldsymbol{p}^{(k)}) \tag{8.3.21}$$

有了 $\boldsymbol{f}^{(k)}$ 后,用(8.3.20)即可求出 $\boldsymbol{r}^{(k)}$,因而,也避免了计算 $\boldsymbol{c}^T\cdot\boldsymbol{c}$。所以,用共轭斜量法解法方程组 $\boldsymbol{c}^T\boldsymbol{cx}+\boldsymbol{c}^T\boldsymbol{d}=0$ 的计算公式可以归纳为:

1) 选择初始向量 $\boldsymbol{x}^{(0)}$,并计算 $\boldsymbol{f}^{(0)}=\boldsymbol{cx}^{(0)}+\boldsymbol{d}$;

2) 对于 $k=1, 2, \cdots$ 计算下列各式:

$$\left.\begin{aligned}
&\boldsymbol{r}^{(k-1)}=\boldsymbol{c}^T\cdot\boldsymbol{f}^{(k-1)}\\
&e_{k-1}=\frac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{r}^{(k-2)}, \boldsymbol{r}^{(k-2)})} \qquad (k\geqslant 2)\\
&\boldsymbol{p}^{(k)}=\begin{cases}-\boldsymbol{r}^{(k-1)} & (k=1)\\ -\boldsymbol{r}^{(k-1)}+e_{k-1}\cdot\boldsymbol{p}^{(k-1)} & (k\geqslant 2)\end{cases}\\
&q_k=\frac{(\boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)})}{(\boldsymbol{cp}^{(k)}, \boldsymbol{cp}^{(k)})}\\
&\boldsymbol{x}^{(k)}=\boldsymbol{x}^{(k-1)}+q_k\cdot\boldsymbol{p}^{(k)}\\
&\boldsymbol{f}^{(k)}=\boldsymbol{f}^{(k-1)}+q_k\cdot\boldsymbol{cp}^{(k)}
\end{aligned}\right\} \tag{8.3.22}$$

上述算法在数学上与对于法方程组直接使用共轭斜量法(8.2.86)是等价的。因而,若不考虑舍入误差的影响,最多计算 m 步就得出真解。此外,还可以证明它有如下重要性质:

定理 3.1 由共轭斜量法(8.3.22)计算而得的误差向量序列$\{\boldsymbol{f}^{(k)}\}$,其模必定是严格地单调减小的。

证明:从(8.3.22)中 $\boldsymbol{f}^{(k)}$ 的递推公式,容易得到:

$$(\boldsymbol{f}^{(k-1)}, \boldsymbol{f}^{(k-1)})=(\boldsymbol{f}^{(k)}, \boldsymbol{f}^{(k)})-2q_k(\boldsymbol{f}^{(k)}, \boldsymbol{cp}^{(k)})+q_k^2(\boldsymbol{cp}^{(k)}, \boldsymbol{cp}^{(k)})$$

注意到 $\boldsymbol{r}^{(k)}$ 与 $\boldsymbol{p}^{(k)}$ 是相互正交的,则有:

$$(\boldsymbol{f}^{(k)}, \boldsymbol{cp}^{(k)})=(\boldsymbol{c}^T\boldsymbol{f}^{(k)}, \boldsymbol{p}^{(k)})=(\boldsymbol{r}^{(k)}, \boldsymbol{p}^{(k)})=0$$

所以,

$$(\boldsymbol{f}^{(k)}, \boldsymbol{f}^{(k)})=(\boldsymbol{f}^{(k-1)}, \boldsymbol{f}^{(k-1)})-q_k^2\cdot(\boldsymbol{cp}^{(k)}, \boldsymbol{cp}^{(k)})$$

只要 $\boldsymbol{x}^{(k-1)}$ 不等于(8.3.5)的真解 $\boldsymbol{x}$,则 $\boldsymbol{r}^{(k-1)}$ 必非零,从而,根据(8.3.22)知 $\boldsymbol{p}^{(k)}$ 亦非零(因 $\boldsymbol{r}^{(k-1)}$ 与 $\boldsymbol{p}^{(k-1)}$ 正交)。再从 $\boldsymbol{c}$ 之秩为 m 得知 $\boldsymbol{cp}^{(k)}\neq 0$,同时 q_k 亦非零。因而,上式说明 $\boldsymbol{f}^{(k)}$ 之模是严格地单调减小的,这就是要证明的结果。

上述定理在实际计算时是有用的。如果按(8.3.22)式计算若干步(少于 m 步)后,误差向量的模$\|\boldsymbol{f}^{(k)}\|$已满足给定的精度要求,那么,根据这个定理,我们就可以停止计算,因为尽管还未求得真解,但若继续计算下去,$\|\boldsymbol{f}^{(k)}\|$不会再超过指定的精度要求,故已不必要了。许

⊖ 这里的 $\boldsymbol{r}^{(k)}$ 与(8.2.86)中的反号。

多实际问题中，往往计算步数比m少得多时就已达到$\|f^{(k)}\|$的精度要求。及时停止计算常常会节省大量的运算时间。

从(8.3.22)式还可看出，共轭斜量法比较适合于处理矩阵C为高阶稀疏矩阵的问题。由于每步计算只需作矩阵C与向量$p^{(k)}$以及矩阵C^T和向量$f^{(k-1)}$的乘积，我们只需将C之全部非零元素及其位置表存放起来(或由程序临时产生这些元素)，就可以简单地完成这些运算。这样，无论存储量及运算量均较节省。同时，由于计算过程中总是与原始矩阵C发生联系，从数值稳定性的角度来看亦是有利的。再加上这个方法也可以看作为一个迭代过程，如果法方程组比较病态，计算m步后可能得不到满意解答，那么，还可以继续计算下去，只要法方程组不是十分病态，一般来说，是可以求得满意结果的。因而可以说，对于高阶稀疏的但不是十分病态的问题，共轭斜量法是较好的方法。本章最后所附的程序16是按照(8.3.22)编制的共轭斜量法解线性矛盾方程组程序。

附录　线性代数方程组的求解程序

一、列主元素消去法解线性代数方程组程序 $GAU(A,\ X,\ N,\ M,\ eps)$

使用说明

本过程是应用列主元素消去法解线性代数方程组：$AX=F$。使用带有回代过程的高斯消去法，并允许带有 M 个右端项。

其中　N——方程组的阶数；

M——右端项的个数；

A——系数矩阵，并按行存放，存放的形式为：

$$a_{11},\ a_{12},\ \cdots,\ a_{1N},\ a_{21},\ a_{22},\ \cdots,\ a_{2N},\ \cdots,\ a_{N1},\ a_{N2},\ \cdots,\ a_{NN};$$

A 的定义为 $A[1:N,\ 1:N]$；

x——右端项矩阵也按行存放，存放的形式为：

$$x_{11},\ x_{12},\ \cdots,\ x_{1M},\ x_{21},\ x_{22},\ \cdots,\ x_{2M},\ \cdots,\ x_{N1},\ x_{N2},\ \cdots,\ x_{NM};$$

x_{ij} 表示第 i 个方程的第 j 个右端项，x 的定义为 $x[1:N,\ 1:M]$；x 又作为存放相应结果的单元。

eps——主元素不能小于的数值，如某个主元素的绝对值小于 eps 则产生停机，继续启动后返回主程序；

程序

```
过程 GAU(A, X, N, M, EPS);
值 N, M, EPS; 场 A, X;
始  简变  B, MAX, JK;
    对于 K=1 到 N 步长 1 执行
    始 0⇒MAX;
      对于 J=K 到 N 步长 1 执行
      若 § ABS(A[J, K])≤MAX 则否始 J⇒JK; § ABS(A[J, K])⇒MAX 终;
      若 MAX<EPS 则始停 11; 转 L 终否;
      1/A[JK, K]⇒MAX;
      对于 J=1 到 N 步长 1 执行
      始 A[JK, J]*MAX⇒B; A[K, J]⇒A[JK, J]; B⇒A[K, J] 终;
      对于 J=1 到 M 步长 1 执行
      始 X[JK, J]*MAX⇒B; X[K, J]⇒X[JK, J]; B⇒X[K, J] 终;
      对于 I=K+1 到 N 步长 1 执行
      始 A[I, K]⇒B;
        对于 J=K+1 到 N 步长 1 执行 A[I, J]−A[K, J]*B⇒A[I, J];
```

```
        对于 J=1 到 M 步长 1 执行 X[I, J]-X[K, J]*B⇒X[I, J]
      终
    终;    注 {消去结束,下面进行回代工作}
    对于 I=N 到 2 步长 1 执行
      对于 J=1 到 M 步长 1 执行
        对于 K=1 到 I-1 步长 1 执行
        X[K, J]-A[K, I]*X[I, J]⇒X[K, J];
      L:
    终;    (注: 过程 GAU 结束)
```

二、全主元素消去法解线性代数方程组程序 $GAUS(A, X, N, M, eps)$

使用说明

本过程是应用全主元素消去法解线性代数方程组 $AX=F$，使用高斯-若当方法，并允许带有 M 个右端项。

其中 N——方程组的阶数；

M——右端项的个数；

A——系数矩阵，并按行存放，存放的形式为：

$$a_{11},\ a_{12},\ \cdots,\ a_{1N},\ a_{21},\ a_{22},\ \cdots,\ a_{2N},\ \cdots,\ a_{N1},\ a_{N2},\ \cdots,\ a_{NN};$$

A 的定义为 $A[1:N, 1:N]$；

x——右端项矩阵也按行存放，存放的形式为：

$$x_{11},\ x_{12},\ \cdots,\ x_{1M},\ x_{21},\ x_{22},\ \cdots,\ x_{2M},\ \cdots,\ x_{N1},\ x_{N2},\ \cdots,\ x_{NM};$$

x_{ij} 表示第 i 个方程的第 j 个右端项，x 的定义为 $x[1:N, 1:M]$；x 又作为最后存放求得的相应结果的单元。

eps——主元素不能小于的数值，如某个主元素的绝对值小于 eps 则产生停机，继续启动后返回主程序；

程序

```
过程 GAUS(A, X, N, M, EPS);
值 N, M, EPS; 场 A, X;
始  简变  B, MAX, IK, JK;
    对于 K=1 到 N 步长 1 执行
    始 0⇒MAX;
        对于 I=K 到 N 步长 1 执行
          对于 J=1 到 N 步长 1 执行
          若 §ABS(A[I, J])≤MAX 则否始 I⇒IK; J⇒JK; §ABS(A[I, J])⇒MAX 终;
        若 MAX<EPS 则始停 11; 转 L 终否;
        1/A[IK, JK]⇒MAX;
        对于 J=1 到 N 步长 1 执行
```

```
始 A[JK, J]*MAX⇒B; A[K, J]⇒A[IK, J]; B⇒A[K, J]终;
对于 J=1 到 M 步长 1 执行
始 X[IK, J]*MAX⇒B; X[K, J]⇒X[IK, J]; B⇒X[K, J]终;
对于 I=1 到 N 步长 1 执行
若 I=K 则 否 始 A[I, JK]⇒B;
                对于 J=1 到 N 步长 1 执行 A[I, J]-A[K, J]*B⇒A[I, J];
                对于 J=1 到 M 步长 1 执行 X[I, J]-X[K, J]*B⇒X[I, J]
            终
终; 注 {求出方程组的解，下面对结果进行排列，排列的原则是按照系数矩阵中 1 出现的位置来进行}
对于 I=1 到 N-1 步长 1 执行
始 对于 K=I 到 N 步长 1 执行
     若 § ABS(A[K, I])<0.5 则
     否始若 K=I 则
          否始对于 J=I+1 到 N 步长 1 执行
               A[I, J]⇒A[K, J];
               对于 J=1 到 M 步长 1 执行
               始 X[K, J]⇒B; X[I, J]⇒X[K, J]; B⇒X[I, J]终
            终; 转 L2
        终;
    L2:
终;
L:
终; 注 {过程 GAUS 结束}
```

三、直接分解法解线性代数方程组程序

UNSYM(A, B, N, R, C, INT, eps)

使用说明

本过程使用 *L-U* 分解法解线性代数方程组 $\boldsymbol{AX}=\boldsymbol{B}$。它分成两个部分：首先把系数矩阵 $\boldsymbol{A}$ 分解为一个下三角矩阵 $\boldsymbol{L}$ 和一个主对角线为 1 的上三角矩阵 $\boldsymbol{U}$ 之积；$\boldsymbol{A}=\boldsymbol{L}\cdot\boldsymbol{U}$。第二部分求解 $\boldsymbol{LUX}=\boldsymbol{B}$。

其中 N——方程组的阶数；

R——右端项的个数；

A——系数矩阵，并按行存放，存放的形式为：

$$a_{11},\ a_{12},\ \cdots,\ a_{1N},\ a_{21},\ a_{22},\ \cdots,\ a_{2N},\ \cdots,\ a_{N1},\ a_{N2},\ \cdots,\ a_{NN};$$

A 的定义为 $A[1:N,\ 1:N]$；以后 A 存放 L 及 U。

B——右端项矩阵，也按行存放，存放的形式为：

$$b_{11},\ b_{12},\ \cdots,\ b_{1R},\ b_{21},\ b_{22},\ \cdots,\ b_{2R},\ \cdots,\ b_{N1},\ b_{N2},\ \cdots,\ b_{NR};$$

b_{ij} 表示第 i 个方程的第 j 个右端项，B 的定义为：$B[1:N,\ 1:R]$；B 最后存放相应的结果。

INT——一个 N 个单元的数组，作为工作单元使用，它存放选取的主元所在行号的序列；

C——使用本过程的标记，当 $C=0$ 时，则本过程的两部分内容都要进行运算；而当 $C=1$ 时，则表示 A 矩阵已分解为 L 及 U，并仍存放在 A 中，*INT* 中也产生了必要的信息，即已经二次以上使用本过程，仅仅换了右端项矩阵。可直接解方程组 $\boldsymbol{LUX}=\boldsymbol{B}$，从而节省了时间。

eps——主元素不能小于的数值，如果主元素的绝对值小于 *eps*，则产生停机，继续启动后返回主程序；

程序

```
过程 UNSYM(A, B, N, R, C, INT, EPS);
场 A, B, INT; 值 N, R, C, EPS;
始  简变 X, Y, L;
    若 C=0 则否转 C1;
    对于 I=1 到 N 步长 1 执行
    始 0⇒Y; 对于 J=1 到 N 步长 1 执行 Y+A[I, J]*A[I, J]⇒Y; 1/§SQRT(Y)⇒
      INT[I]
    终;
    对于 K=1 到 N 步长 1 执行
    始 K⇒L; 0⇒X;
      对于 I=K 到 N 步长 1 执行
      始 -A[I, K]⇒Y;
        对于 J=1 到 K-1 步长 1 执行 Y+A[I, J]*A[J, K]⇒Y; -Y⇒A[I, K];
          §ABS(Y*INT[I])⇒Y;
        若 X<Y 则始 Y⇒X; I⇒L 终否
      终;
      若 L=K 则否始对于 J=1 到 N 步长 1 执行
                    始 A[K, J]⇒Y; A[L, J]⇒A[K, J]; Y⇒A[L, J] 终;
                    INT[K]⇒INT[L]
                  终;
      L⇒INT[K];
      若 X<8*EPS 则始停 11; 转FAIL 终否; 1/A[K, K]⇒X;
      对于 J=K+1 到 N 步长 1 执行
      始 -A[K, J]⇒Y;
        对于 I=1 到 K-1 步长 1 执行 Y+A[K, I]*A[I, J]⇒Y; X*Y⇒A[K, J]
      终
    终; 注{本过程第一部分分解完毕,下面是第二部分}
  C1: 对于 K=1 到 N 步长 1 执行
    若 INT[K]=K 则
        否  对于 J=1 到 R 步长 1 执行
```

```
        始 B[K, J]⇒Y; B[INT[K], J]⇒B[K, J];
          Y⇒B[INT[K], J]
        终;
  对于 K=1 到 R 步长 1 执行
  始对于 I=1 到 N 步长 1 执行
    始 B[I, K]⇒Y; 对于 J=1 到 I-1 步长 1 执行 Y+A[I, J]*B[J, K]⇒Y; Y/A
      [I, I]⇒B[I, K]
    终;
    对于 I=N 到 1 步长 -1 执行
    始 B[I, K]⇒Y;
      对于 J=I+1 到 N 步长 1 执行 Y+A[I, J]*B[J, K]⇒Y; -Y⇒B[I, K]
    终
  终;
 FAIL:
终; 注 {过程 UNSYM 结束}
```

四、平方根法解对称正定线性代数方程组程序 $LLTS(n, R, S\theta L, A, B, D1, D2)$

使用说明

本程序是用平方根法求解对称正定线性代数方程组，可以同时处理 R 个($R\geqslant1$) 自由项，求解过程中附带算出系数矩阵的行列式值 $\det A=2^{d_2}\times d_1$。如果使用本程序之前，系数矩阵已进行过分解，可令参数 $S\theta L=1$，程序自动跳过分解步骤去解相应的三角型方程组。如果在舍入误差范围内系数矩阵是奇异的，程序执行过程中将出现停机(停机号码为 555)。

输入参数：

n——方程组的阶数。

A——存放系数矩阵上三角部分元素的一维场，其定义为 $A[1:n*(n+1)/2]$。元素按行存放，其次序为：$a_{11}, a_{12}, \cdots, a_{1n}, a_{22}, \cdots, a_{2n}, a_{33}, \cdots, a_{nn}$。

R——自由项的列数。

B——存放自由项矩阵(其每一列为一个自由项)的元素的二维场，其定义为 $B[1:n, 1:R]$。自由项元素按行存放，其次序为：$b_{11}, b_{12}, \cdots, b_{1r}, b_{21}, b_{22}, \cdots, b_{2r}, \cdots, b_{n1}, \cdots, b_{nr}$。

$S\theta L$——若 $S\theta L=0$，表示系数矩阵需要进行分解。若 $S\theta L=1$，表示系数矩阵已进行过分解，并将上三角因子 L^T 的元素按行存于场 A。程序执行过程中将自动跳过相应分解步骤。

输出参数：

$D2$——存系数矩阵行列式值的二进制比例因子的阶码。

$D1$——存系数矩阵行列式引入比例因子后的值，即 $\det A=2^{d_2}\times d_1$。

B——程序工作完毕后，存放方程组的解答 x。其存放次序与自由项相同。

程序

```
过程 LLTS(N, R, SθL, A, B, D1, D2);
  值 N, R, SθL; 场 A, B;
  简变 D1, D2;
    始  简变 S, X;
      若 SθL=1 则转 SL 否;
      1⇒D1; 0⇒D2;
      对于 I=1 到 N 步长 1 执行
        对于 J=I 到 N 步长 1 执行
          始 0⇒S; 0⇒X;
            对于 K=1 到 I-1 步长 1 执行
            始 X+A[I+S]*A[J+S]⇒X; S+N-K⇒S
            终;
            A[J+S]-X⇒X;
            若 I=J 则
            始若 X≤0 则停 555 否;
              D1*X⇒D1;
            SN: 若 1<D1 则
                    始 D1*0.125⇒D1;
                      D2+3⇒D2;
                      转 SN
                    终
                        否
            MN: 若 D1<0.125 则
                      始 D1*8⇒D1;
                        D2-3⇒D2;
                        转 MN
                      终
                                否;
            1/§SQRT(X)⇒A[I+S]
                终  否
            X*A[I+S]⇒A[J+S]
            终;
  SL: 对于 J=1 到 R 步长 1 执行
        始 对于 I=1 到 N 步长 1 执行
          始 0⇒S; B[I, J]⇒X;
            对于 K=1 到 I-1 步长 1 执行
            始 X-A[I+S]*B[K, J]⇒X; S+N-K⇒S
            终;
```

```
            X*A[I+S]⇒B[I, J];
              I+S⇒S
          终;
          对于 I=N 到 1 步长 -1 执行
            始 B[I, J]⇒X;
              对于 K=N 到 I+1 步长 -1 执行
              始 X-A[S]*B[K, J]⇒X; S-1⇒S
              终;
              X*A[S]⇒B[I, J];
              S-1⇒S
            终
        终
    终;
```

五、LDL^T 分解法解对称正定线性代数方程组程序 $LDLT(n, R, S\theta L, A, B, D1, D2)$

使用说明

本程序是用 LDL^T 分解法解系数矩阵为对称正定矩阵的线性代数方程组。计算公式见公式(8.1.26)。本程序除假定系数矩阵 $\boldsymbol{A}$ 的下三角部分以按行排列的次序存于场 $\boldsymbol{A}$ 外，其它规定完全与程序四相同，故这里不再说明。矩阵 $\boldsymbol{A}$ 的元素存于场 $\boldsymbol{A}$ 中的次序为：a_{11}，a_{21}，a_{22}，a_{31}，a_{32}，a_{33}，a_{41}，…，a_{n1}，a_{n2}，…，a_{nn}。所以，场 A 的定义仍为 $\boldsymbol{A}[1:n*(n+1)/2]$。

程序

```
过程 LDLT(N, R, SθL, A, B, D1, D2);
  值 N, R, SθL; 场 A, B; 简变 D1, D2;
  始  简变 X, IK, JK, Y, Z;
      若 SθL=0 则转 SθLE 否;
      1⇒D1; 0⇒D2;
      对于 I=1 到 N 步长 1 执行
      始 I*(I-1)/2⇒IK;
        对于 J=1 到 I-1 步长 1 执行
        始 J*(J-1)/2⇒JK; A[IK+J]⇒X;
          对于 K=1 到 J-1 步长 1 执行
          X-A[IK+K]*A[JK+K]⇒X; X⇒A[IK+J]
        终;
        A[IK+I]⇒X;
        对于 K=1 到 I-1 步长 1 执行
        始 A[IK+K]⇒Y; Y/A[K*(K-1)/2]⇒Z; Z⇒A[IK+K]; X-Y*Z⇒X
        终;
```

```
        X⇒A[IK+I]; D1*X⇒D1;
        若 X=0 则始 0⇒D2; 停 555 终否;
     L1: 若 1≤§ABS(D1)则始 D1*0.125⇒D1;
                          D2+3⇒D2; 转 L1
                            终
                          否;
     L2: 若 §ABS(D1)<0.1 则始 D1*8⇒D1;
                          D2-3⇒D2; 转 L2
                            终
                          否
     终;
  SOLE:
     对于 I=1 到 N 步长 1 执行
     始 I*(I-1)/2⇒IK;
          对于 Q=1 到 R 步长 1 执行
          对于 K=1 到 I-1 步长 1 执行
          B[I, Q]-A[IK+K]*B[K, Q]⇒B[I, Q]
     终;
     对于 I=1 到 N 步长 1 执行
     始 I*(I-1)/2⇒IK;
       对于 Q=1 到 R 步长 1 执行 B[I, Q]/A[IK+K]⇒B[I, Q]
     终;
     对于 I=N 到 1 步长 -1 执行
     始 I*(I-1)/2⇒IK;
          对于 Q=1 到 R 步长 1 执行
          对于 K=1 到 I-1 步长 1 执行
          B[K, Q]-A[IK+K]*B[I, Q]⇒B[K, Q]
     终
   终;
```

六、解线性代数方程组的镜像映射法程序 $H\Theta US(A, B, X, EPS, IN, m, n)$

使用说明

使用者在套用本程序前，必须先在主程序中定义场 A、B、x，三个场分别用来存放系数阵 $\boldsymbol{A}$，右端矩阵 $\boldsymbol{B}$ 及解矩阵 $\boldsymbol{X}$，都是按列存放。

形式参数 IN 用作标志。IN 为正值时程序执行三角化过程；IN 为负值时程序跳过三角化部分而直接进入回代求解。IN 要按需要给以正值或负值，0 作正值处理。

形式参数 EPS 的值为机器表示的最小正数。当程序中发生 $SM<EPS$ 时，系数阵 $\boldsymbol{A}$

在舍入误差范围内是奇异的，将出现程序停机。

使用者需自编印刷语句打印解 x。

形式参数

输入参数：

A——$A[1:n, 1:n]$，用来存放系数阵的场，按列存放。

B——$B[1:m, 1:n]$，存放右端矩阵的场，按列存放。

EPS——机器表示的最小正数。

IN——作标志用，$IN<0$ 时不作三角化直接求解，否则作三角化并求解。

n——方程组的阶数。

m——右端向量个数。

输出参数：

x——$x[1:m, 1:n]$，存放方程组解的场，按列存放。

程序

```
过程 HθUS(A, B, X, EPS, IN, N, M);
    场 A, B, X;
    值 EPS, IN, N, M;
    始  简变 ALJ, AQJ, BETA, SM;
        场 Y, AD[1:N];
        若 IN<0 则转 SΘLVE 否;
        对于 J=1 到 N 步长 1 执行
        始 0⇒SM;
          对于 I=J 到 N 步长 1 执行
          A[J, I]↑2+SM⇒SM;
          若 SM<EPS 则停否;
 注 {当 SM<EPS 时，A 为奇异，程序停机}
          A[J, J]⇒AQJ;
          若 AQJ<0 则 §SQRT(SM)⇒ALJ
                   否 §SQRT(SM)⇒ALJ;
          ALJ⇒AD[J];
          1/(SM-AQJ*ALJ)⇒BETA;
          AQJ-ALJ⇒A[J, J];
          若 J<N 则
          始 对于 K=J+1 到 N 步长 1 执行
            始 0⇒SM;
              对于 I=J 到 N 步长 1 执行
              A[J, I]*A[K, I]+SM⇒SM; BETA*SM⇒Y[K]·
              对于 I=J 到 N 步长 1 执行
              A[K, I]-Y[K]*A[J, I]⇒A[K, I]
            终
```

```
        终否;
注 {上面对第J+1列到第N列作变换。下面对右端矩阵B作同样变换}
        对于 K=1 到 M 步长 1 执行
        始 0⇒SM;
          对于 I=J 到 N 步长 1 执行
          A[J, I]*B[K, I]+SM⇒SM;
          BETA*SM⇒SM;
          对于 I=J 到 N 步长 1 执行
          B[K, I]-SM*A[J, I]⇒B[K, I]
        终
      终;
SOLVE:
        对于 I=N 到 1 步长 -1 执行
          对于 J=1 到 M 步长 1 执行
          始 B[J, I]/AD[I]⇒B[J, I];
            B[J, I]⇒X[J, I];
            若 1<I 则
            对于 K=I 到 N 步长 1 执行
            B[J, I-1]-A[K, I-1]*B[J, K]⇒B[J, I-1]
            否
          终
    终;
```

七、对称正定矩阵原地求逆程序
GJINVER(n, A)

使用说明

本程序是用高斯-若当顺序消去法对于对称正定矩阵进行原地求逆。求逆过程中每一中间矩阵均是对称的，所以只存储其上三角部分元素。如果在舍入误差范围内矩阵是奇异的，程序执行过程中将出现停机(停机号码为555)。

参数表:

n——矩阵的阶数。

A——存放矩阵上三角部分元素的一维场，其定义为 $A[1:n*(n+1)/2]$。元素按行存放，其次序为: $a_{11}, a_{12}, a_{13}, \cdots, a_{1n}, a_{22}, a_{23}, \cdots, a_{2n}, a_{33}, \cdots, a_{nn}$。程序工作完毕后，场 A 中按上述同样顺序存放逆矩阵上三角部分的元素。

程序

```
过程 GJINVER(N, A);
  场 A; 值 N;
  始  简变 P, Q, PP, QQ; 场 H[1:N];
```

```
对于 K=N 到 1 步长 -1 执行
始 A[1]⇒P;
    若 P⩽0 则停 555 否;
    对于 I=2 到 N 步长 1 执行
    始 A[I]⇒Q;
      若 K<I 则 Q⇒QQ 否 -Q⇒QQ; QQ/P⇒H[I]; 0⇒PP;
      对于 J=2 到 I 步长 1 执行
        始 PP+N-J+1⇒PP; A[I+PP]+Q*H[J]⇒A[I+PP-N+J-2]
        终
    终;
    1/P⇒A[N+PP];
    对于 J=1 到 N-1 步长 1 执行
    始 PP-J⇒PP; H[N-J+1]⇒A[N+PP] 终
终
终;
```

八、全主元素消去法求逆矩阵程序

GASI(*A*, *N*, *eps*)

使用说明

本过程是应用全主元素消去法求矩阵 A 的逆矩阵，使用了 2*N 个工作单元，是一种原地求逆的方法。

其中 N——矩阵的阶数；

A——被求逆的矩阵，并按行存放，存放的形式为：

$$a_{11},\ a_{12},\ \cdots,\ a_{1N},\ a_{21},\ a_{22},\ \cdots,\ a_{2N},\ \cdots,\ a_{N1},\ a_{N2},\ \cdots,\ a_{NN};$$

A 的定义形式为 $A[1:N,\ 1:N]$；A 最后存放逆矩阵。

eps——主元素不能小于的数值，如果主元素的绝对值小于 *eps*，则产生停机，继续启动后返回主程序；

程序

```
过程 GASI(A, N, EPS);
  场 A; 值 N, EPS;
  始  简变 MAX, B, IK, JK; 场 Z[1:2*N];
     过程 F(K); 值 K;
     始若 IK=K 则否对于 J=1 到 N 步长 1 执行
                  始 A[IK, J]⇒B; A[K, J]⇒A[IK, J]; B⇒A[K, J] 终;
       若 JK=K 则否对于 J=1 到 N 步长 1 执行
                  始 A[J, JK]⇒B; A[J, K]⇒A[J, JK]; B⇒A[J, K] 终
     终;
     对于 K=1 到 N 步长 1 执行
```

```
始 0⇒MAX;
  对于 I=K 到 N 步长 1 执行
    对于 J=K 到 N 步长 1 执行
    若 §ABS(A[I, J])≤MAX 则否始 I⇒IK; J⇒JK; §ABS(A[I, J])⇒MAX 终;
  若 MAX<EPS 则始停 11; 转 L 终否;
  1/A[IK, JK]⇒MAX; 1⇒A[IK, JK];
  IK⇒Z[2*K-1]; JK⇒Z[2*K]; F(K);
  对于 J=1 到 N 步长 1 执行 A[K, J]*MAX⇒A[K, J];
  对于 I=1 到 N 步长 1 执行
  若 I=K 则否始 A[I, K]⇒MAX; 0⇒A[I, K];
              对于 J=1 到 N 步长 1 执行 A[I, J]-MAX*A[K, J]⇒A[I, J]
              终
终; 注 {已求出逆矩阵,下一步进行排列}
对于 K=N-1 到 1 步长 -1 执行
始 Z[2*K]⇒IK; Z[2*K-1]⇒JK; F(K)终;
L:
终; 注 {过程 GAS 结束}
```

九、平方根法解带型对称正定线性代数方程组程序 $BDLT(n, r, m, S\theta L, A, B, d1, d2)$

使用说明

本程序是用平方根法求解带型对称正定线性代数方程组。由于系数矩阵的带型特点，本程序只要求存放系数矩阵上三角部分处于非零元素带以内的元素。也可以同时处理 R 个 ($R\geqslant1$) 自由项，求解过程中并附带算出系数矩阵的行列式值 $\det \boldsymbol{A}=2^{d2}\times d1$。如果使用本程序之前，系数矩阵已进行过分解(例如，对于同一系数矩阵第二次使用本程序)，可令参数 $S\theta L=1$，程序自动跳过分解步骤去解相应的三角型方程组。如果在舍入误差范围内系数矩阵是奇异的，程序执行过程中将出现停机(停机号码为555)。

输入参数:

n——方程组的阶数。

m——系数矩阵的"半带宽"(带宽为 $2m+1$)。

A——存放系数矩阵上三角部分的非零元素带的二维场，其定义为: $A[1:n, 0:m]$。元素按行存放，其次序为: $a_{11}, a_{12}, \cdots, a_{1,m+1}, a_{22}, a_{23}, \cdots, a_{2,m+2}, \cdots, a_{33}, \cdots, a_{3,m+3}, \cdots, a_{nn}, \underbrace{x, x, \cdots, x}_{m\text{个}}$。详见式(8.1.40)。

R——自由项的列数。

B——存放自由项矩阵(其每一列为一个自由项)的元素的二维场，其定义为: $B[1:n, 1:R]$。自由项元素按行存放，其次序为: $b_{11}, b_{12}, \cdots, b_{1R}, \cdots, b_{21}, b_{22}, \cdots, b_{2R}, \cdots, b_{n1}, \cdots, b_{nR}$。

$S\theta L$——若 $S\theta L=0$ 表示系数矩阵需要进行分解，若 $S\theta L=1$，表示系数矩阵已进行过分解，并将上三角因子 L^T 的元素按与原始矩阵同样方式存于场 A。程序执行过程中将自动跳过相应分解步骤。

输出参数:

$D2$——存放系数矩阵行列式值的二进制比例因子的阶码。

$D1$——存放系数矩阵行列式引入比例因子后的值，即 $\det \boldsymbol{A}=2^{d2}\times \boldsymbol{d1}$。

B——程序工作完毕后，存放方程组的解答 X，其存放次序与自由项相同。

程序

```
过程 BDLT(N, R, M, SθL, A, B, D1, D2);
  值 N, R, M, SθL; 场 A, B;
  简变 D1, D2;
  始  简变 S, L, X;
      若 SθL=1 则转 SL 否;
      1⇒D1; 0⇒D2;
      对于 I=1 到 N 步长 1 执行
      始若 I⩽N-M 则 M⇒S 否 N-I⇒S;
        对于 J=0 到 S 步长 1 执行
          始若 I+J⩽M 则 I-1⇒L 否 M-J⇒L; A[I, J]⇒X;
            对于 K=1 到 L 步长 1 执行 X-A[I-K, K]*A[I-K, K+J]⇒X;
            若 J=0 则
              始若 X⩽0 则停 555 否;
                D1*X⇒D1;
                SN: 若 1<D1 则始 D1*0.125⇒D1; D2+3⇒D2; 转 SN
                               终
                              否;
                MN: 若 D1<0.125 则始 D1*8⇒D1; D2-3⇒D2; 转 MN
                                    终
                                   否;
                1/§SQRT(X)⇒A[I, 0]
              终
                否
              X*A[I, 0]⇒A[I, J]
          终
      终;
   SL: 对于 J=1 到 R 步长 1 执行
        始  对于 I=1 到 N 步长 1 执行
          始 B[I, J]⇒X;
            若 I⩽M 则 I-1⇒S 否 M⇒S;
            对于 K=1 到 S 步长 1 执行
```

```
      X-A[I-K, K]*B[I-K, J]⇒X; X*A[I, 0]⇒B[I, J]
    终;
    对于 I=N 到 1 步长 -1 执行
      始 B[I, J]⇒X;
        若 I≤N-M 则 M⇒S 否 N-I⇒S;
        对于 K=1 到 S 步长 1 执行
          X-A[I, K]*B[I+K, J]⇒X; X*A[I, 0]⇒B[I, J]
      终
    终
终;
```

十、变带宽对称正定线性方程组求解程序 *SLDLT*(*n*, *R*, *S*, *A*, *B*, *DA*)

使用说明

本程序是用 LDL^T 分解法求解系数矩阵为变带宽型对称正定矩阵的线性方程组，可以同时处理 R 个($R\geqslant1$)自由项。如果使用本程序前，系数矩阵 $\boldsymbol{A}$ 已进行过分解，并按要求方式存于场 A 中，则可令 $S=0$，程序执行过程中自动跳过相应的分解步骤。

为了节省存储单元，本程序采取紧缩存储方式。即将矩阵 $\boldsymbol{A}$ 下三角部分的每一行中，从第一个非零元素开始至对角线元素为止的所有元素依次存放于一维场 A 中。场 A 的定义为：$A[1:NS]$，其中 NS 为所需存储的元素总个数。为了标明各行元素在场 A 中的位置，另外用一个一维场 DA 记录各行对角线元在场 A 中的位置，其定义为 $DA[0:n]$。例如，若矩阵 $\boldsymbol{A}$ 为下列形式：

$$\boldsymbol{A}=\begin{bmatrix} 1 & & & & & & \\ 0.2 & 2 & & & \text{对} & \text{称} & \\ 0 & 0.1 & 3 & & & & \\ 0 & 0.3 & 0.5 & 4 & & & \\ 0 & 0.2 & 0 & 0.4 & 5 & & \\ 0 & 0 & 0 & 0 & 0 & 6 & \\ 0 & 0 & 0.6 & 0 & 0 & 1 & 7 \end{bmatrix}$$

则场 A 中存放的元素为：1, 0.2, 2, 0.1, 3, 0.3, 0.5, 4, 0.2, 0, 0.4, 5, 6, 0.6, 0, 0, 1, 7。场 DA 的内容为：0, 1, 3, 5, 8, 12, 13, 18(为方便起见，我们假定 $DA[0]\equiv0$)。按照这一存储格式，容易看出矩阵 $\boldsymbol{A}$ 的 i 行 j 列元素在场 A 中的位置为：$DA[i]-i+j$。

如果用 N_i 表示矩阵 $\boldsymbol{A}$ 第 i 行第一个非零元素的列号，根据式(8.1.26)可以推知此时有如下计算公式：

$$\boldsymbol{A}=\boldsymbol{L}\cdot\boldsymbol{D}\cdot\boldsymbol{L}^T$$

$\boldsymbol{L}=[l_{ij}]$ 为单位下三角形矩阵

$\boldsymbol{D}=\mathrm{diag}(d_i)$

其中元素 l_{ij}, d_i 的计算公式如下：

令 $ji=\max(N_i, N_j)$; $\tilde{a}_{ij}=l_{ij}d_j$。

$$\begin{cases}\tilde{a}_{ij}=a_{ij}-\sum\limits_{k=ji}^{j-1}\tilde{a}_{ik}\cdot l_{jk}, \quad l_{ij}=\tilde{a}_{ij}/d_j \quad (j=N_i, N_{i+1}, \cdots, i-1)\\ d_i=a_{ii}-\sum\limits_{k=N_i}^{i-1}\tilde{a}_{ik}\cdot l_{ik}\end{cases}$$

$(i=1, 2, \cdots, n)$

求得元素 l_{ij} 及 d_i 后，再解如下方程组即得最终解答 X：

$$\begin{cases}\boldsymbol{L}\cdot\boldsymbol{Z}=\boldsymbol{B}\\ \boldsymbol{D}\cdot\boldsymbol{Y}=\boldsymbol{Z}\\ \boldsymbol{L}^T\cdot\boldsymbol{X}=\boldsymbol{Y}\end{cases}$$

形式参数

输入参数：

n——方程组的阶数。

R——自由项的列数。

A——存放矩阵 $\boldsymbol{A}$ 下三角部分元素的一维场，其定义为：$A[1:NS]$，其中 NS 为需存放元素的总数。元素存放的方式见前面的使用说明。本程序工作完毕后，场 A 中按同样方式存放着下三角因子 L 和对角阵 D 的元素。D 的元素处于原来矩阵 $\boldsymbol{A}$ 对角线元位置，L 的元素(指 $i>j$ 的部分)存于矩阵 $\boldsymbol{A}$ 的相应元素位置上。

DA——存放矩阵 $\boldsymbol{A}$ 对角线元素在场 A 中位置的一维场，其定义为：$DA[0:n]$。本程序要求 $DA[0]$ 存以零。从 $DA[1]$ 起，依次存放各行对角线元素的位置(见前面的例)。

B——按行存放自由项矩阵(其每一列为一个自由项)元素的二维场，其定义为：$B[1:n, 1:R]$。元素存放次序为：$b_{11}, b_{12}, \cdots, b_{1r}, b_{21}, \cdots, b_{2r}, \cdots, b_{n1}, \cdots, b_{nr}$。

S——若 $S=0$，表示矩阵 $\boldsymbol{A}$ 已进行过分解，并已将 L 及 D 的元素按前述规定存于场 A，程序执行过程中自动跳过相应的分解步骤。否则，程序执行分解矩阵 $\boldsymbol{A}$ 的步骤。

输出参数

B——本程序工作完毕后，场 B 中存放着要求的解答 X，其存放方式与自由项相同。

程序

```
过程 SLDLT(N, R, S, A, B, DA);
  值 N, R, S;
  场 A, B, DA;
  始  简变 NI, NJ, IJ, JI, IK, JK, Y, Z;
      若 S=0 则转 SθLE 否;
      对于 I=1 到 N 步长 1 执行
      始 I-DA[I]+DA[I-1]+1⇒NI; DA[I]-I⇒IK;
        对于 J=NI 到 I 步长 1 执行
```

```
始 J－I+DA[I]⇒IJ; J－DA[J]+DA[J－1]+1⇒NJ;
    若 NI≤NJ 则 NJ⇒JI 否 NI⇒JI; DA[J]－J⇒JK;
    对于 K=JI 到 J－1 步长 1 执行
      始 A[IK+K]⇒Y;
        若 J=I 则始 A[JK+K]/A[DA[K]]⇒Z; Z⇒A[JK+K]终
               否 A[JK+K]⇒Z;
                      A[IJ]－Y*Z⇒A[IJ]
      终
  终
 终;
SθLE:
 对于 I=1 到 N 步长 1 执行
 始 I－DA[I]+DA[I－1]+1⇒NI; DA[I]－I⇒IK;
   对于 Q=1 到 R 步长 1 执行
   对于 K=NI 到 I－1 步长 1 执行
   B[I, Q]－A[IK+K]*B[K, Q]⇒B[I, Q]
 终;
 对于 I=1 到 N 步长 1 执行
 对于 Q=1 到 R 步长 1 执行 B[I, Q]/A[DA[I]]⇒B[I, Q];
 对于 I=N 到 1 步长 －1 执行
 始 I－DA[I]+DA[I－1]+1⇒NI; DA[I]－I⇒IK;
   对于 Q=1 到 R 步长 1 执行
   对于 K=NI 到 I－1 步长 1 执行
     B[K, Q]－A[IK+K]*B[I, Q]⇒B[K, Q]
 终
终;
```

十一、追赶法解三对角线方程组程序
TRDIAG (*A*, *X*, *N*)

使用说明

本过程是应用追赶法求解三对角线的方程组 $\boldsymbol{AX}=\boldsymbol{F}$。其中 N 是方程的阶数；$\boldsymbol{A}$ 为系数矩阵

$$\boldsymbol{A}=\begin{pmatrix} b_1 & c_1 & & & \mathbf{0} \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & \mathbf{0} & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix}$$

$\boldsymbol{A}$ 按对角线存放：$a_1, a_2, \cdots, a_n, b_1, b_2, \cdots, b_n, c_1, c_2, \cdots, c_n$；并令 $a_1=c_n=0$。A 的定义形式为 $A[1:3, 1:N]$；场 X 为存放右端项 $\boldsymbol{F}$ 用，又作为存放方程组的解 $\boldsymbol{X}$ 用，其定义形式为 $X[1:N]$；

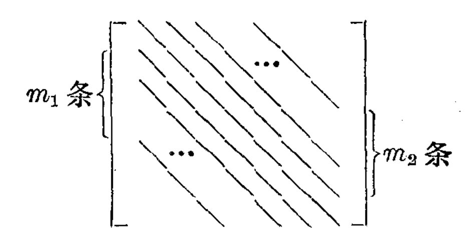
程序

```
过程 TRDIAG(A, X, N);
  场 A, X; 值 N;
  始  对于 I=1 到 N-1 步长 1 执行
      始 -A[3, I]/A[2, I]⇒A[3, I];  A[3, I]*A[1, I+1]+A[2, I+1]⇒A[2, I+1]
      终;
      X[1]/A[2, 1]⇒X[1];
      对于 I=1 到 N-1 步长 1 执行
      (X[I+1]-X[I]*A[1, I+1])/A[2, I+1]⇒X[I+1];
      对于 I=N-1 到 1 步长 -1 执行
      X[I+1]*A[3, I]+X[I]⇒X[I]
  终;
```

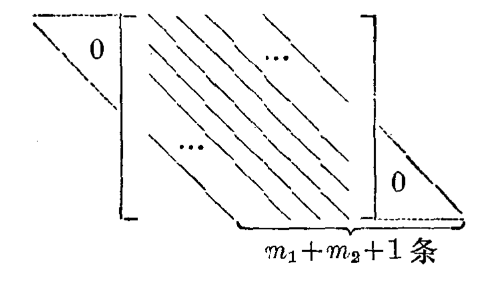
十二、列主元素法解非对称带状方程组程序 $UNSB(A, B, EPS, n, r, \mathrm{m}_1, m_2)$

使用说明

本程序适用于求解非对称带状方程组 $\tilde{\boldsymbol{A}}\boldsymbol{x}=\tilde{\boldsymbol{B}}$，其中 $\tilde{\boldsymbol{A}}$ 是非对称带状矩阵，即当 $i>j+m_1$ 或 $j>i+m_2$ 时 $\tilde{a}_{ij}=0$。$\tilde{\boldsymbol{A}}$ 的形状为



系数阵 $\tilde{\boldsymbol{A}}$ 存放在一个 $n\times(m_1+m_2+1)$ 的场 A 中，A 定义为 $[1:n, -m_1:m_2]$，使用者先把矩阵 $\tilde{\boldsymbol{A}}$ 的左上角和右下角补 0 使成为如下形状：



然后按行存放在场 A 中。

$\tilde{\boldsymbol{B}}$ 是右端矩阵，存于场 B 中，也是按行存放。在套用本程序前使用者还需在主程序中定义简变 EPS，在用过程语句前必须对 EPS 赋值，其值为机器表示的最小正数，当程序中

出现 $x<EPS$ 时，表示考虑到舍入误差的影响，方程组的系数阵奇异，此时本程序不能求解，出现程序停机。

作好上述准备后，使用者可用前面的过程语句求解，求解后须自编印刷语句打印结果，解答存放在场 B 中。

形式参数表：

输入参数：

A——$A[1:n,\ -m_1:m_2]$，存放系数矩阵，存放方式如说明所述。

B——$B[1:n,\ 1:r]$，存放右端矩阵，按行存放。

n——方程组的阶数。

m_1——对角线下面的条数，即下带宽。

m_2——对角线上面的条数，即上带宽。

EPS——机器表示的最小正数。

输出参数：

B——$B[1:n,\ 1:r]$，求解后存放解矩阵。

程序

```
过程 UNSB(A, B, EPS, N, R, M1, M2);
  场 A, B;
  简变 EPS;
  值 N, R, M1, M2;
  始  简变 X, L, T;
      M1⇒L;
      对于 I=1 到 M1 步长 1 执行
      始对于 J=1-I 到 M2 步长 1 执行
        A[I,J]⇒A[I, J-L]; L-1⇒L;
        对于 J=M2-L 到 M2 步长 1 执行
        0⇒A[I, J]
      终;
注：{上面这段程序把 A 的前 M1 行元素左移，右边补 0}
      M1⇒L;
      对于 K=1 到 N 步长 1 执行
      始 A[K, -M1]⇒X; K⇒T;
        若 L<N 则 L+1⇒L 否;
        对于 J=K+1 到 L 步长 1 执行
        若 §ABS(X)<§ABS(A[J, -M1])则
        始 A[J, -M1]⇒X; J⇒T
        终  否;
注  {选列主元，X 中为主元，T 中是主元对应的行号}
        若 §ABS(X)<EPS 则停否;
注  {§ABS(X)<EPS 时，在舍入误差范围内方程组的系数阵奇异，程序停机}
```

```
        若 T=K 则否
        始  对于 J=-M1 到 M2 步长 1 执行
          始 A[K, J]⇒X; A[T, J]⇒A[K, J]; X⇒A[T, J]
          终;
          对于 J=1 到 R 步长 1 执行
          始 B[K, J]⇒X; B[T, J]⇒B[K, J]; X⇒B[T, J]
          终
        终;
注 {进行行交换}
        对于 I=K+1 到 L 步长 1 执行
        始 A[I, -M1]/A[K, -M1]⇒X;
          对于 J=1-M1 到 M2 步长 1 执行
          A[I, J]-X*A[K, J]⇒A[I, J-1]; 0⇒A[I, M2];
          对于 J=1 到 R 步长 1 执行
          B[I, J]-X*B[K, J]⇒B[I, J]
        终
注 {执行消去步骤,下面是回代求解}
      终;
SOLVE:
      对于 J=1 到 R 步长 1 执行
      始 -M1⇒L;
        对于 I=N 到 1 步长 -1 执行
        始 B[I, J]⇒X; I+M1⇒T;
          对于 K=1-M1 到 L 步长 1 执行
          X-A[I, K]*B[K+T, J]⇒X; X/A[I, -M1]⇒B[I, J];
          若 L<M2 则 L+1⇒L 否
        终
      终
    终;
```

十三、共轭斜量法解线性代数方程组程序

$CGM(A, U, B, n, eps)$

使用说明

本程序是用共轭斜量法解线性代数方程组 $\boldsymbol{AU}=\boldsymbol{B}$,其中,$\boldsymbol{A}$ 是 n 阶对称正定矩阵; $\boldsymbol{U}$, $\boldsymbol{B}$ 是 n 维向量。为了克服计算中容易发生溢出的缺点(R、P 的欧氏模平方都容易上溢、下溢),本程序引入一个比例因子,使得修正向量 $\boldsymbol{P}$ 的绝对值最大的分量接近于1。假定初始向量为零,算法可写为下列形式:

(1) $d_1=2^w\approx\max\limits_i(|b_i|)$; 其中 w 是整数。

$$\boldsymbol{P}^{(1)}=\boldsymbol{R}^{(0)}=\boldsymbol{B}/d_1$$

(2) 对于 $k=1, 2, 3, \cdots$ 执行

$$q_k=(R^{(k-1)}, R^{(k-1)})/(AP^{(k)}, P^{(k)})$$

$$q_k=q_k d_k$$

$$\boldsymbol{U}^{(k)}=\boldsymbol{U}^{(k-1)}+q_k\boldsymbol{P}^{(k)}$$

$$\boldsymbol{R}^{(k)}=\boldsymbol{R}^{(k-1)}-q_k\boldsymbol{AP}^{(k)}$$

$$e_k=(\boldsymbol{R}^{(k)}, \boldsymbol{R}^{(k)})/(\boldsymbol{R}^{(k-1)}, \boldsymbol{R}^{(k-1)})$$

$$\boldsymbol{P}^{(k+1)}=\boldsymbol{R}^{(k)}+e_k\boldsymbol{P}^{(k)}$$

$$\bar{d}_{k+1}=2^w\approx\max_i(|\bar{p}_i^{(k+1)}|)$$

$$\boldsymbol{P}^{(k+1)}=\boldsymbol{P}^{(k+1)}/\bar{d}_{k+1}$$

$$\boldsymbol{R}^{(k)}=\boldsymbol{R}^{(k)}/\bar{d}_{k+1}$$

$$d_{k+1}=d_k\bar{d}_{k+1}$$

计算直到下列条件之一满足后停止。其中条件(3)表示计算结果不能满足要求。

1) $(\boldsymbol{R}, \boldsymbol{R})\leqslant eps$,

2) $(\boldsymbol{AP}, \boldsymbol{P})\leqslant eps$;

3) 迭代次数 $=3n$

计算结果存储在场 U 中

形式参数表

A——场 $A[1:n*(n+1)/2]$，系数矩阵，按上三角形存放为：$a_{11}, a_{12}, \cdots a_{1n}, a_{22}, a_{23}, \cdots, a_{2n}, \cdots, a_{nn}$。

B——场 $B[1:n]$ 常数项，按次序排为 $b_1, b_2, \cdots, b_n$。

n——值 n 方程组阶数。

eps——值 eps 方程剩余的平方和 (R, R) 或 (AP, P) 的允许误差。

$iter$——迭代次数。

U——场 $U[1:n]$ 方程组的解。

程序

```
过程 MA(A, X, Y);
场 A, X, Y;
  始  简变 H, W;
     对于 I=1 到 N 步长 1 执行
     始(I-1)*N-(I-1)*I/2⇒W; 0⇒H;
       对于 J=1 到 N 步长 1 执行
         若 I≤J 则 H+A[W+J]*X[J]⇒H
         否 H+A[(J-1)*N+I-(J-1)*J/2]*X[J]⇒H; H⇒Y[I]
     终
  终;
```

注 {过程 MA 是提供给过程 CGM 作 Y=AX 运算用的，其中 A 按上三角形存储。根据 A 的特殊形状，使用者可另行改编}

```
过程 CGM(A, U, B, N, EPS);
```

```
值 N, EPS;
场 A, U, B;
始  简变 Q, Q1, Q2, E, E1, W, ITER, D, G, EPS;
    场 S, R, P, AP[1:N];
  0⇒D;
  对于 I=1 到 N 步长 1 执行
  始  §ABS(B[I])⇒G;
    若 G⩽D 则否 G⇒D
  终;
    §LN(D)/0.693⇒W;
    §ENTI(W)⇒W; 2↑W⇒D;
  对于 I=1 到 N 步长 1 执行
    B[I]/D⇒R[I]; R⇒P; 0⇒Q1; 0⇒U;
  对于 I=1 到 N 步长 1 执行
    Q1+R[I]↑2⇒Q1;
  对于 K=1 到 3*N 步长 1 执行
  注  {迭代开始,并限定次数不大于 3N}
  始 EP/D/D⇒EPS; MA(A, P, AP); 0⇒Q2;
    对于 I=1 到 N 步长 1 执行
    Q2+AP[I]*P[I]⇒Q2;
    若 Q2<0 则停 123 否;
    若 Q2<EPS 则转 END 否;
    Q1/Q2⇒Q; 0⇒E1; Q*D⇒G;
    对于 I=1 到 N 步长 1 执行
    始 U[I]+G*P[I]⇒U[I]; R[I]-Q*AP[I]⇒W; W⇒R[I]; E1+W*W⇒E1
    终;
    若 E1<EPS 则转 END 否;
    E1/Q1⇒E; E1⇒Q1;
    对于 I=1 到 N 步长 1 执行
    R[I]+E*P[I]⇒P[I]; D⇒E; 0⇒D;
    对于 I=1 到 N 步长 1 执行
    始 §ABS(P[I])⇒G;
      若 G⩽D 则否 G⇒D
    终;
        §LN(D)/0.693⇒W;
        §ENTI(W)⇒W; 2↑W⇒D;
    对于 I=1 到 N 步长 1 执行
    始 P[I]/D⇒P[I]; R[I]/D⇒R[I]
    终;
```

```
    Q1/D/D⇒Q1; D*E⇒D;  注 {假定小于10⁻⁹的数看作0}
    若 D≤(10↑(-9))则转 END 否; K⇒ITER
 终;
 END: 印+ITER
终;
```

十四、解线性矛盾方程组的镜像映射法程序

$H\Theta US(A, B, X, EPS, RD, m, n)$

使用说明

使用者在套用本程序前，必须先在主程序中定义场 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{X}$ 和简变 EPS、RD，场 $\boldsymbol{A}$ 用来存放系数矩阵 $\boldsymbol{A}$，其元素按列存放。即以 a_{11}, a_{21}, $\cdots$, a_{m1}, a_{12}, $\cdots$, a_{mn} 的次序存放。简变 EPS 需由使用者在调用本程序前赋值，其值为机器表示的最小正数。

计算结果放在场 x 中，RD 中是余量模的平方，即 $\|\boldsymbol{B}-\boldsymbol{AX}\|_2^2$。

当程序中发生 $SM<EPS$ 时，系数阵的计算秩小于 n，此时方程没有唯一的最小二乘解，程序停机。

使用者需在主程序中自编印刷语句，打印解 x 及余量 RD。

形式参数表:

输入参数

A——$A[1:n, 1:m]$，存放方程组的系数阵的场，按列存放。

B——$B[1:m]$，存放右端向量的场。

EPS——机器表示的最小正数。

m——方程的个数。

n——未知数的个数。

输出参数

x——$x[1:n]$，存放解向量的场。

RD——存放余量模的平方。

程序

```
过程 HΘUS(A, B, X, EPS, RD, M, N);
    场 A, B, X;
    简变 EPS, RD;
    值 M, N;
    始  简变 ALJ, AQJ, BETA, SM;
        场 Y, AD[1:N];
        对于 J=1 到 N 步长 1 执行
        始 0⇒SM;
          对于 I=J 到 M 步长 1 执行
          A[J, I]↑2+SM⇒SM;
          若 SM<EPS 则停否;
```

注 {简变 SM 中为 σ_j^2, 当 SM<EPS 时, A 的计算秩小于 N, 此时方程没有唯一解,程序停机。}

```
        A[J, J]⇒AQJ;
        若 AQJ<0 则 §SQRT(SM)⇒ALJ
                 否 §SQRT(SM)⇒ALJ;
        ALJ⇒AD[J];
        1/(SM－AQJ*ALJ)⇒BETA;
        AQJ－ALJ⇒A[J, J];
        若 J<N 则
        始  对于 K=J+1 到 N 步长 1 执行
          始 0⇒SM;
            对于 I=J 到 M 步长 1 执行
            A[J, I]*A[K, I]+SM⇒SM; BETA*SM⇒Y[K];
            对于 I=J 到 M 步长 1 执行
            A[K, I]－Y[K]*A[J, I]⇒A[K, I]
          终
        终
              否;
```

注 {上面 K 循环是对第 J+1 列到第N列作 HOUSEHOLDER 变换}

```
        0⇒SM;
        对于 I=J 到 M 步长 1 执行
        A[J, I]*B[I]+SM⇒SM;
        BETA*SM⇒SM;
        对于 I=J 到 M 步长 1 执行
        B[I]－A[J, I]*SM⇒B[I]
```

注 {对右端 B 作同样的变换}

```
      终;
```

注 {三角化步骤已完成,下面用回代过程求解 x}

```
      对于 I=N 到 1 步长 －1 执行
      始 B[I]/AD[I]⇒B[I]; B[I]⇒X[I];
        若 1<I 则
          对于 K=I 到 N 步长 1 执行
          B[I－1]－A[K, I－1]*B[K]⇒B[I－1]
                否
      终;
```

注 {求解结束,下面计算余量}

```
      0⇒RD;
      对于 I=N+1 到 M 步长 1 执行
      B[I]↑2+RD⇒RD
    终;
```

十五、解线性矛盾方程组的正交化法程序 $MGS(A, B, X, p, EPS, m, n)$

使用说明

使用者在套用本程序前，必须先在主程序中定义场 **A**、**B**、**X**、**P**，场 **A**、**B** 分别存放系数矩阵和右端向量，系数阵 **A** 按列存放在场 **A** 中。

本程序带选主元的，以未正交化的各列中模最大的列作为主列，场 **P** 中先是存放各列模的平方，后存放对应的主列的列号，作恢复解分量的次序用。

在程序中 $MAX<EPS$ 时，系数阵 **A** 的计算秩小于 n，方程组无唯一的最小二乘解，将出现程序停机。

使用者需自编印刷语句，打印结果。

形式参数 EPS 的值为机器表示的最小正数。

形式参数表:

输入参数

A——$A[1:n, 1:m]$，存放方程组的系数阵，按列存放。

B——$B[1:m]$，存放右端向量的场。

m——方程的个数。

n——未知数的个数。

p——$p[1:n]$，先存放各列模的平方，后存放主列的列号，作恢复解分量次序用。

EPS——机器表示的最小正数。

输出参数

x——$x[1:n]$，存放解向量。

程序

```
过程 MGS(A, B, X, P, EPS, M, N);
    场 A, B, X, P;
    值 EPS, M, N;
    始  简变 S, T, MAX;
        对于 I=1 到 N 步长 1 执行
        始 0⇒S;
          对于 J=1 到 M 步长 1 执行
          A[I, J]↑2+S⇒S; S⇒P[I]
        终;
  注 {上面计算各列的模的平方，以便选主元}
        对于 I=1 到 N 步长 1 执行
        始 0⇒MAX;
          对于 J=I 到 N 步长 1 执行
          若 MAX<P[J] 则
          始 P[J]⇒MAX; J⇒T 终
```

```
      否;
      若 MAX<EPS 则停否;
注 {MAX 中为主元列的模的平方，当 MAX<EPS 时，方程没有唯一解，程序停机}
      P[I]⇒P[T];
      T⇒P[I];
      对于 J=1 到 M 步长 1 执行
      始 A[I, J]⇒S; A[T, J]⇒A[I, J]; S⇒A[T, J]
      终;
注 {上面是选出主元列后，把有关的列的元素及信息进行交换}
      MAX⇒X[I];
      对于 K=I+1 到 N 步长 1 执行
      始 0⇒S;
        对于 J=1 到 M 步长 1 执行
        A[I, J]*A[K, J]+S⇒S; S⇒X[K]; S/MAX⇒T;
        对于 J=1 到 M 步长 1 执行
        A[K, J]-A[I, J]*T⇒A[K, J]; P[K]-T*S⇒P[K]
      终;
注 {上面对第 I+1 列到第 N 列作正交化，同时计算各列的模平方}
      0⇒S;
      对于 J=1 到 M 步长 1 执行
      A[I, J]*B[J]+S⇒S;
      对于 K=I 到 N 步长 1 执行
      X[K]⇒A[I, K]; S⇒X[I]
    终;
注 {正交化步骤完成，以下是回代求解 X}
SOLVE.
    对于 I=N 到 1 步长 -1 执行
    始 X[I]⇒S;
      对于 J=I+1 到 N 步长 1 执行
      S-A[I, J]*X[J]⇒S;
      S/A[I, I]⇒X[I];
      P[I]⇒T;
      X[I]⇒S;
      X[T]⇒X[I];
      S⇒X[T]
注 {后四个语句是按选主元列时的信息恢复解 X 的原始次序}
    终
  终;
```

十六、共轭斜量法解线性矛盾方程组程序

$CGV(m, n, c, x, d)$

使用说明

本程序是用共轭斜量法计算矛盾方程组

$$CX+D=F$$

的最小二乘解 X。其中 C 是 $m\times n$ 阶矩阵 D 是 m 维向量，X 是 n 维向量。若 C 的秩 $<n$ 则计算将停机。

形式参数表：

c——场 $c[1:m, 1:n]$ 按行排列为：c_{11}, c_{12}, …, c_{1n}, c_{21}, c_{22}, …, c_{2n}, …, c_{m1}, c_{m2}, …, c_{mn}。

d——场 $d[1:m]$ 方程常数项。

m——值 m, c 的行数。

n——值 n, c 的列数。

eps——值 eps, 是 x 的允许误差。

x——场 $x[1:n]$，解向量。

$iter$——迭代次数。

程序

```
过程 MCX(C, X, Y);
场 C, X, Y;
始  0⇒Y;
    对于 I=1 到 M 步长 1 执行
    对于 J=1 到 N 步长 1 执行
    Y[I]+C[I, J]*X[J]⇒Y[I]
终;
    注 {过程 MCX 是提供给过程 CGV 计算 Y=CX 的}
过程 MCTE(C, F, R);
场 C, F, R;
始  0⇒R;
    对于 J=1 到 N 步长 1 执行
    对于 I=1 到 M 步长 1 执行
      R[J]+C[I, J]*F[I]⇒R[J]
终;
    注 {本过程是提供给 CGV 程序计算 R=CᵀF 用的}
过程 CGV(M, N, EPS, C, X, D);
值 M, N, EPS;
场 C, X, D;
  始简变 Q, Q1, Q2, E1, E2, E, RE1, RF2, ITER;
```

```
场 S, F, R[1:N], P, CP[1:M]; D⇒F; 0⇒RE2; 0⇒X;
对于 K=1 到 3*N 步长 1 执行
注 {限制迭代次数小于 3N+1 次}
始 K⇒ITER; MCTF(C, F, R); 0⇒E1;
对于 I=1 到 N 步长 1 执行
    E1+R[I]↑2⇒E1;
若 K=1 则
始 对于 I=1 到 N 步长 1 执行
-R[I]⇒P[I]; 转 L1
终
 否;
若 E2<(10↑(-9))则转END 否;
注 {假定小于 10⁻⁹ 的数都看作零}
      E1/E2⇒E;
对于 I=1 到 N 步长 1 执行
-R[I]+E*P[I]⇒P[I];
L1:E1⇒E2; MCX(C, P, CP); 0⇒Q2;
对于 I=1 到 M 步长 1 执行
Q2+CP[I]↑2⇒Q2;
若 Q2<0 则停 123 否;
若 Q2<(10↑(-9))则转 END
否;
E1/Q1⇒Q;
对于 I=1 到 N 步长 1 执行
X[I]+Q*P[I]⇒X[I]; 0⇒RE1;
对于 J=1 到 M 步长 1 执行
始 F[J]+Q*CP[J]⇒F[J]; RE1+F[J]↑2⇒RE1
终;
若 §ABS(RE1-RE2)≤§ABS(RE1*EPS)则转 END 否;
RE1⇒RE2
终;
END: 印 +ITER;
终;
```

参 考 资 料

[1] 中国科学院计算技术研究所编,《计算方法讲义》,科学出版社, 1958。

[2] 北京大学等编,《计算方法》,人民教育出版社, 1961。

[3] A. 拉尔斯登、H S. 维尔夫著,徐献瑜等译,《数字计算机上用的数学方法》,卷 1,上海科学技术出版社, 1963。

[4] Ф. Р. 甘特马赫著,柯召译,《矩阵论》,高等教育出版社, 1955。

[5] В. В. Воеводин, "Численные Методы Алгебры", Москва, 1966.

[6] Д. К. 法捷也夫、B. H. 法捷也娃著,刘光武等译,«线性代数计算方法»,上海科学技术出版社, 1965。

[7] G. E. 福雪斯、W. R. 华沙著,胡祖炽等译,«偏微分方程的有限差分法»,上海科学技术出版社, 1964。

[8] G. E. Forsythe, C. B. Moler, "Computer Solution of linear Algebraic Systems", Prentice-Hall, Englewood Cliffs, N. J. 1967.

[9] L. Fox, "An Introduction to Numerical Linear Algebra", Clarendon, Oxford, 1964.

[10] A. S. Householder, "The Theory of matrices in numerical Analysio", Blaisdell, Newyork, 1964.

[11] A. Ralston, H. S. Wilf, "Mathe matical methods for Digital Computers", Vol. II, John Wiley & Sons, INC. 1967.

[12] Schwarz/Rutishauser/Stiefel, "Matrizen-Numerik", B. G. Teubner, Stuttgart, 1972.

[13] R. S. 瓦格著,蒋尔雄等译,«矩阵迭代分析»,上海科学技术出版社, 1966。

[14] E. L. Wachsprep, "Iterative Solution of Elliptic systems and Applications to the Neutron Cliffusion Equations of reactor physics", Englewood cliffs, 1966.

[15] J. R. Westlake, "A Handbook of Numerical Matrix Inversion and solution of linear Eqautions", John Wiley & Sons, Inc. 1968.

[16] J. H. Wilkinson, "Rounding Errors in Algebraic processes", prentice-Hall, Englewood cliffs, New Jersey 1963.

[17] J. H. Wilkinson, "The Algebraic Eigenvalue problem", Clarendon press, Oxbord, 1965.

[18] J. H. Wilkinson, C. Reinsch, "Handbook for Automatic Computation", Vol. II, Linear Algebra, Berlin-Heidelberg-New york, 1971.

[19] D. M. Young, "Iterative solution of large linear systems", Academic press, New york and London, 1971.

[20] B. A.Carre, "The Determination of the Optimum accelerating factor for successive Over-Relaxation", «The Computer Journal», 4(1961), pp. 73～78.

[21] J. K. Reid, "A method for finding the Optimum successive Over-Relaxation parameter", «The Computer Journal» 9(1966), pp.200～204.

[22] G. H. Golub. J. H. Wilkinson,"Note on the Iterative Refinement of Least Squares Solution",«Num. Math.» Band 9(1966) Heft2 pp. 139～148.

[23] R. S. Anderssen, "Computational Considerations for Linear Least Squares Methods," «Least Squares Methods in Data Analysis», 1969, pp. 19～31.

[24] I. S. Duff, "A Survey of Sparse Matrix Research", IEEE Proceedings, 1977, Vol. 65, No. 4, pp. 500～535.

[25] R. P. Tewarson, "Sparse Matrices" New York: Academic Press, 1973.

[26] W. Niethammer, "Über-und Unterrelaxation bei linearen Gleichungssystemen" 〈Computing〉 Vol.5 (1970) pp. 303～311.

第九章　非线性方程和非线性方程组的解法

§9.1　引　　言

在科学研究和工程设计中常常遇到求解一非线性方程或非线性方程组的问题。例如求解方程

$$x^4-10x^3+35x^2-50x+24=0 \tag{9.1.1}$$

或

$$e^{-x}-\sin\left(\frac{\pi x}{2}\right)=0 \tag{9.1.2}$$

方程(9.1.1)的左端是未知量 x 的多项式，这类方程称为高次代数方程；在方程(9.1.2)中，因左端包含的指数和正弦函数是属超越函数，这样的方程称为超越方程。

下面我们用符号 $f(x)$ 来表示方程左端的函数，于是可写方程的一般形式为

$$f(x)=0 \tag{9.1.3}$$

方程的解通常称为方程的根，或称为函数 $f(x)$ 的零点。

方程的根可能是实数也可能是复数，相应地称为实根和复根。如果对于数 α 有 $f(\alpha)=0$，但 $f'(\alpha)\neq 0$（$f'(x)$ 表示 $f(x)$ 的一阶导数），则 α 称为单根；如果有 $f(\alpha)=f'(\alpha)=\cdots=f^{(k)}(\alpha)=0$ 但 $f^{(k+1)}(\alpha)\neq 0$，则 α 称为 k 重根（此处 $f^{(k)}(\alpha)$ 表示 $f(x)$ 的 k 阶导数）。

对于高次代数方程，其根（实或复的）的个数与其次数相同。如方程(9.1.1)是4次方程，就有4个根。至于超越方程，其解可能是一个或几个，也可能是无穷多个。

常见的求解问题有如下两种要求：一种是要求定出在给定范围内的某个解，而解的粗略位置事先已从问题的物理背景或应用其他方法得知。另一种是定出方程的全部解，或者定出在给定区域内的所有解，而解的个数和位置事先并不知道，这在超越方程的情形是比较困难的。

求解非线性方程，除少数特殊方程（例如二次多项式方程）可以利用公式直接定出它的零点外，一般只能采用某种迭代解法，即从预知的解的初始近似值（简称初值）开始，利用某种迭代格式构造一近似值序列 $x_0, x_1, \cdots, x_k, x_{k+1}, \cdots$，逐步逼近于所求的解案 α。

这个序列是否确能收敛于解，以及是否很快地逼近于所求的解案，这是迭代解法的收敛性和收敛速度的问题。

对于一种解法，为了考察它的有效性，一般都要讨论它的收敛性和收敛速度，即考察在什么样的条件下构造的序列是收敛的，以及序列中的近似值又按什么样的误差下降速度逼近解。

迭代过程的收敛条件，一般与方程的性态（函数 $f(x)$ 在解案附近的性质，零点的分布状况等）以及初值的近似度有关。某些解法仅与初值的近似度有关，故有时亦称收敛条件为收敛范围。

迭代过程的收敛速度，是指在接近收敛的过程中近似值误差的下降速度。一般说来，它主要由方法所决定。方程的性态也会起一些影响。

在本章中，我们将介绍几种对两类方程均适用的较有效的方法。对于高次代数方程尚有不少特殊的有效方法，此处就不介绍了。在所介绍的方法中，除了区间分半法仅限于求实根外，其他均无此限制。这几种方法一次都只求出一个解。因此，我们将在§9.7中结合二次插值法和线性分式插值法介绍关于求方程若干个解或在指定范围内所有解的处理方法。

§9.2 求实根的区间分半法

9.2.1 方法简述

作为一个例子，让我们求解超越方程

$$e^{-x}-\sin\left(\frac{\pi x}{2}\right)=0 \tag{9.2.1}$$

这个方程在0与1之间有一个解，要求将解定到小数点后两位的精确度。

记$f(x)\equiv e^{-x}-\sin\left(\frac{\pi x}{2}\right)$，这个函数是连续的。我们知道，对于连续函数$f(x)$，如果在$x=a$和$x=b$处的量值$f(a)$和$f(b)$有相异的符号(简称异号)，那末根据$f(x)$的连续性，在区间$[a, b]$上至少存在$f(x)$的一个零点。据此，可以采用如下的过程来定出解的精确位置。

先算出$f(x)$在区间$[0, 1]$端点处的量值:

$$f(0)=1;\quad f(1)=e^{-1}-\sin\frac{\pi}{2}\approx-0.632121$$

然后将区间$[0, 1]$分半，分成二个小区间:

$$[0, 0.5];\quad [0.5, 1]$$

计算$f(x)$在小区间端点$x=0.5$处的量值:

$$f(0.5)=-0.100576$$

因为$f(0)$与$f(0.5)$异号，而$f(0.5)$与$f(1)$同号，因此知解位于$[0, 0.5]$的小区间内。于是定解区间的长度缩小了一半。然后，再对小区间$[0, 0.5]$按上述过程进行。如此继续下去，区间一次次缩半，直到区间长度缩到0.01以下时为止。其具体过程见表9.1。

表9.1 用区间分半法定方程$e^{-x}-\sin\left(\frac{\pi x}{2}\right)=0$的解

x	$f(x)$	存在解的区间
0	1	
1	−0.632121	[0, 1]
0.5	−0.100576	[0, 0.5]
0.25	0.396117	[0.25, 0.5]
0.375	0.131719	[0.375, 0.5]
0.4375	0.011255	[0.4375, 0.5]
0.46875	−0.045775	[0.4375, 0.46875]
0.453125	−0.017534	[0.4375, 0.453125]
0.4453125	−0.003208	[0.4375, 0.4453125]
0.44140625	即作为所求的解	

上面结合例子介绍了区间分半法，为清楚起见，再将区间分半法的一般执行步骤归纳如下：

9.2.2 执行步骤

(1) 计算$f(x)$在区间$[a, b]$(存在解的)端点处的值$f(a)$, $f(b)$;

(2) 计算$f(x)$在区间中点$\frac{a+b}{2}$处的值$f\left(\frac{a+b}{2}\right)$;

(3) 判断：若$f\left(\frac{a+b}{2}\right)=0$则$\frac{a+b}{2}$即是解，否则检验：

若$f\left(\frac{a+b}{2}\right)$与$f(a)$异号，则知解位于区间$\left[a, \frac{a+b}{2}\right]$中，以$\frac{a+b}{2}$代替$b$;

若$f\left(\frac{a+b}{2}\right)$与$f(b)$异号，则知解位于区间$\left[\frac{a+b}{2}, b\right]$中，以$\frac{a+b}{2}$代替$a$。

反复执行步骤(2)、(3)，直到区间长度缩小到允许误差范围之内，此时区间的中点即可作为所要求的解。

上面介绍的区间分半法，每迭代一次，区间缩小一半，也就是说，解的不定范围每次只缩小一半。因此，它是一种收敛比较缓慢的方法，并且，它只能用于求实函数的实零点。下面再介绍几种收敛比较快的并且适用范围比较广的解法，这些方法都基于构造某种插值函数的思想。

§9.3 线性插值法(弦位法)

9.3.1 方法简述

设已知方程

$$f(z)=0 \quad (z=x+iy) \tag{9.3.1}$$

的解α的两个近似值z_1、z_2，设它们相应的函数$f(z)$的值分别是f_1、f_2，即

$$f_1=f(z_1), \quad f_2=f(z_2)$$

通过这两个点

$$(z_1, f_1), (z_2, f_2) \tag{9.3.2}$$

构造一线性函数

$$L(z)=az+b \tag{9.3.3}$$

易于定出此线性函数为

$$L(z)=f_2+\frac{f_2-f_1}{z_2-z_1}(z-z_2) \tag{9.3.4}$$

事实上，将点列(9.3.2)分别代入关系式(9.3.3)，得到二元线性联立方程组

$$\left.\begin{aligned} f_1=az_1+b \\ f_2=az_2+b \end{aligned}\right\}$$

由此解出a、b，便得(9.3.4)。

当z为实变量x且$f(x)$为实函数时，线性函数$L(x)$的几何意义就是通过函数$f(x)$图形上的二点(x, f_1)、(x_2, f_2)所作的弦线，见图9.1。

在近似值z_1、z_2附近，以线性函数$L(z)$来近似函数$f(z)$，且以$L(z)$的零点，记为z_3,

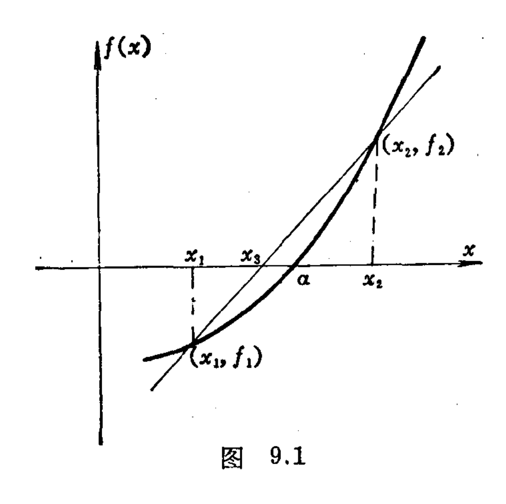


图 9.1

$$z_3=z_2-f_2\Big/\frac{f_2-f_1}{z_2-z_1} \tag{9.3.5}$$

作为解 α 的一个新的近似值(在几何上,就是以通过曲线上二点(x_1, f_1)、(x_2, f_2)的弦线来近似曲线,且以弦线与 x 轴的交点 x_3 来进一步逼近曲线与 x 轴的交点 α,见图 9.1)。

预期 z_3 比 z_1、z_2 更接近于解 α,于是将新的近似值 z_3 代替 z_1。重复上述过程,又得到 z_4,如此反复执行,直到充分逼近于解 α 时为止。

线性插值法,通常亦称弦位法。其收敛性和收敛速度的讨论可详见[1],下面仅引其结论。

9.3.2 方法的收敛性

如果函数 $f(z)$ 在零点 α 附近存在连续的二阶导数,且初始近似值 z_1, z_2 充分接近于所求的零点,那末线性插值法的迭代过程是收敛的。其收敛速度为

$$\lim_{k\to\infty}\frac{|z_{k+1}-\alpha|}{|z_k-\alpha|^{1.618}}=K$$

或者说,当 z_k 充分接近于 α 时有

$$|z_{k+1}-\alpha|\approx K|z_k-\alpha|^{1.618} \tag{9.3.6}$$

其中,z_k 表示迭代过程中的第 k 次近似值;z_{k+1} 表示第 $k+1$ 次近似值;

$$K=\left|\frac{f''(\alpha)}{2f'(\alpha)}\right|^{0.618}$$

$f'(\alpha)$、$f''(\alpha)$分别表示 $f(z)$ 在零点 α 处的一阶、二阶导数值。

关系式(9.3.6)反映了在接近收敛的过程中,用线性插值法每迭代一次近似值误差的缩减速度,或者说有效数字的增长速度。例如,假设

$$\frac{|z_k-\alpha|}{|\alpha|}=0.0498$$

即 z_k 具有二位有效数字,那末从式(9.3.6)有:

$$\frac{|z_{k+1}-\alpha|}{|\alpha|}\approx K0.0498^{1.618}\approx 0.008K\quad\left(K=\left|\frac{f''(\alpha)}{2f'(\alpha)}\right|^{0.618}\right)$$

如果 K 的数值不大,那末 z_{k+1} 便有接近三位的有效数字。由此看出,如果系数 K 的影响可忽略,那末近似值误差的下降速度主要取决于式(9.3.6)中的幂次 1.618。这个幂次可以作为衡量收敛速度的一种标志数,一般称为收敛速度的阶。

一般地,如果由一种迭代解法构造出的近似值序列 z_0, z_1, …, z_k 与解 α 的误差有:

$$\lim_{k\to\infty}\frac{|z_{k+1}-\alpha|}{|z_k-\alpha|^{r}}=K$$

或者,当 z_k 充分接近于解 α 时有关系式

$$|z_{k+1}-\alpha|\approx K|z_k-\alpha|^{r} \tag{9.3.7}$$

其中 $r\geqslant 1$,则称该法具有 r 阶的收敛速度。

一般说来,对于具有 r 阶收敛速度的算法,当接近收敛时其近似值的误差将按幂次为 r

的速度下降，因此 r 越大，误差就下降得越多，收敛速度就越高；r 越小，误差就下降得越少，收敛速度就越低。

对于 $r=1$ 的算法，其收敛性与收敛速度就取决于因子 K。当 $K<1$ 时，其误差将按数 K 比例地下降，称为线性收敛速度。

式(9.3.7)中的因子 K 是与函数在零点附近的性状有关的一个量。它对收敛速度亦会起一定的影响。例如在上述例子中，如果 $K=1$，则在迭代一次后有

$$\frac{|z_{k+1}-\alpha|}{|\alpha|}\approx 0.008$$

如果 $K=5$，则为

$$\frac{|z_{k+1}-\alpha|}{|\alpha|}\approx 0.04$$

与第 k 次近似值 z_k 的误差

$$\frac{|z_k-\alpha|}{|\alpha|}=0.0498$$

相比，新的近似值 z_{k+1} 在 $K=1$ 时增加了有效数位，在 $K=5$ 时便改进不大。因此与函数性状有关的 K 对收敛速度亦有一定的影响。从实际计算亦表明，对于同一种算法，用于不同性态的函数其收敛速度亦存有差异，特别在 r 不大时此种现象更为显著。事实上从改写式(9.3.7)为

$$|z_{k+1}-\alpha|\approx|K^{\frac{1}{r}}(z_k-\alpha)|^r$$

后可看出，如果 K 的量值较大或阶 r 不高使 $K^{\frac{1}{r}}$ 比 1 大得多时，其收敛速度就要慢一些，如果阶 r 较高且 K 的量值不太大，有 $K^{\frac{1}{r}}\approx 1$，则 K 的影响将较小。此外，K 值的大小还影响收敛范围。事实上从反复利用式(9.3.7)可导出关系式

$$|z_k-\alpha|\approx K^{\frac{r^k-1}{r-1}}|z_0-\alpha|^{r^k}=K^{-\frac{1}{r-1}}|K^{\frac{1}{r-1}}(z_0-\alpha)|^{r^k}$$

此处假定了 z_0 已充分接近于 α。如果选择初值 z_0 使有

$$|z_0-\alpha|K^{\frac{1}{r-1}}<1$$

即

$$|z_0-\alpha|<1/K^{\frac{1}{r-1}}$$

则当 $k\to\infty$ 时有 $z_k\to 0$。由此看出，当 K 值越大，要求 z_0 越接近于 α。从实际计算亦表明，一种算法当用于不同性态的函数，对初值近似度的要求亦不同。

9.3.3 计算步骤

(1) 准备：选定初始近似值 z_1、z_2，计算相应的函数值 $f_1=f(z_1)$，$f_2=f(z_2)$。

(2) 迭代：按公式

$$z_3=z_2-f_2\Big/\frac{f_2-f_1}{z_2-z_1}$$

迭代一次，得新的近似值 z_3，计算 $f_3=f(z_3)$。

(3) 控制：如果 z_3 满足 $|\delta|\leqslant\varepsilon$ 或 $|f_3|\sim 0$，则认为过程收敛，终止迭代，以 z_3 作为所求的解案，否则执行步(4)，此处

ε 是允许误差；

$$\delta=\begin{cases}|z_3-z_2|, & \text{当 } |z_3|<c \text{ 时}\\ \dfrac{|z_3-z_2|}{|z_3|}, & \text{当 } |z_3|\geqslant c \text{ 时}\end{cases} \tag{9.3.8}$$

其中 c 是取绝对或相对误差的控制数，一般可取 $c=1$。

(4) 迭代准备：如果迭代次数超过某个上界 $\max k$，则认为过程不收敛，计算失败，否则以 (z_2, f_2)、(z_3, f_3) 分别代替 (z_1, f_1)、(z_2, f_2)，而后转步骤(2)继续迭代。

§9.4 牛 顿 法

9.4.1 方法简述

设已知 z_1 是方程

$$f(z)=0$$

之解 α 的一个近似值。假设函数 $f(z)$ 在零点 α 附近导数存在。设 f_1、f_1' 分别是 $f(z)$ 在 z_1 处的函数和导数值，即 $f_1=f(z_1)$，$f_1'=f'(z_1)$。

构造一线性函数 $L(z)$，使其在 z_1 处取值 f_1，且其斜率为 f_1'。易于导出这样的线性函数是

$$L(z)=f_1'(z-z_1)+f_1 \tag{9.4.1}$$

在 z_1 附近，就以此线性函数 $L(z)$ 来近似函数 $f(z)$，且以 $L(z)$ 的零点，记为 z_2，

$$z_2=z_1-\frac{f_1}{f_1'} \tag{9.4.2}$$

作为解 α 的一个新的近似值。然后以 z_2 代替 z_1，重复上述过程，得到 z_3，如此等等，直到近似值充分接近解案时为止。

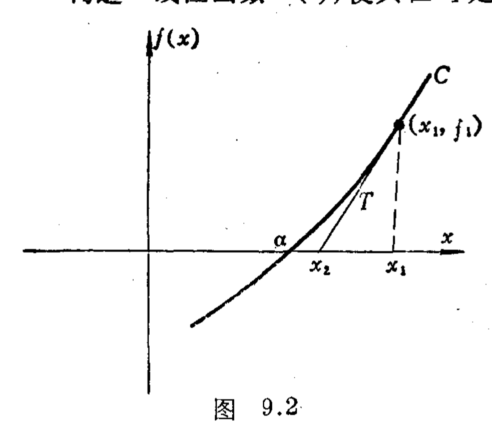


图 9.2

当 z 为实变量 x，$f(x)$ 为实函数时，线性函数 $L(x)$ 代表 $f(x)$ 的曲线 C 在点 (x_1, f_1) 处的切线，见图 9.2。牛顿法的几何意义就是：在近似值 x_1 附近，以曲线在点 (x_1, f_1) 处的切线 T 来近似曲线 C，且以切线与 x 轴的交点作为解 α 的进一步近似。如图 9.2 所示。

9.4.2 方法的收敛性

如果 $f(z)$ 在零点附近存在连续的二阶导数，且初始近似值 z_1 充分接近于所求的解，那末牛顿迭代过程是收敛的，且其收敛速度是：在解 α 为单根的情形，有

$$\lim_{k\to\infty}\frac{|z_{k+1}-\alpha|}{|z_k-\alpha|^2}=\left|\frac{f''(\alpha)}{2f'(\alpha)}\right| \tag{9.4.3}$$

或者说，当 z_k 充分接近于 α 时有

$$|z_{k+1}-\alpha|\approx\left|\frac{f''(\alpha)}{2f'(\alpha)}\right||z_k-\alpha|^2 \tag{9.4.4}$$

这表明，牛顿迭代法具有阶为 2 的收敛速度。此即在迭代接近收敛的过程中，近似值 z_k 的误差将是平方地减小，故又称为“平方收敛速度”。因此，是一种收敛比较快的方法。

在解 α 为 P 重根的情形，有

$$\lim_{k\to\infty}\frac{|z_{k+1}-\alpha|}{|z_k-\alpha|}=\frac{P-1}{P}$$

或者说，当 z_k 充分接近于 α 时，有

$$|z_{k+1}-\alpha|\approx\frac{P-1}{P}|z_k-\alpha| \tag{9.4.5}$$

这表明，在迭代接近收敛的过程中，近似值 z_k 的误差按 $\frac{P-1}{P}$ 的比例线性地下降，为线性收敛速度，因此牛顿法遇重根时收敛速度大大下降，且重数 P 越高，收敛越慢。收敛性的证明可参考[1]，此处从略。

9.4.3 计算步骤

(1) 准备：选定初始近似值 z_1，计算 $f_1=f(z_1)$，$f'_1=f'(z_1)$。

(2) 迭代：按公式

$$z_2=z_1-\frac{f_1}{f'_1}$$

迭代一次，得新近似值 z_2，计算 $f_2=f(z_2)$，$f'_2=f'(z_2)$。

(3) 控制：如果 z_2 满足 $|\delta|<\varepsilon$ 或 $f_2\approx0$，则过程收敛，终止迭代，以 z_2 作为所要求的解案。否则转步(4)。此处，δ，ε 的意义同于(9.3.8)中所述。

(4) 迭代准备：如果迭代次数超过上界或者 $f'_2=0$，则方法失败。否则以 (z_2, f_2, f'_2) 代替 (z_1, f_1, f'_1)，转步 2 继续迭代。

§9.5 二次插值法(Müller 法)

9.5.1 方法简述

设已知非线性方程

$$f(z)=0$$

解的三个近似值 z_1、z_2、z_3，设相应的 $f(z)$ 的函数值是 f_1、f_2、f_3。二次插值法的基本思想是，通过这三个近似值点：

$$(z_1, f_1),\ (z_2, f_2),\ (z_3, f_3) \tag{9.5.1}$$

构造二次函数

$$L(z)=Az^2+Bz+C$$

(当 z 为实数时，此函数图形即是通过上述三点的一条抛物线，见图 9.3)。在近似值附近就以此二次函数 $L(z)$ 来近似非线性函数 $f(z)$，并以二次函数的两零点之一作为解的进一步近似。下面，我们推导它的计算公式。

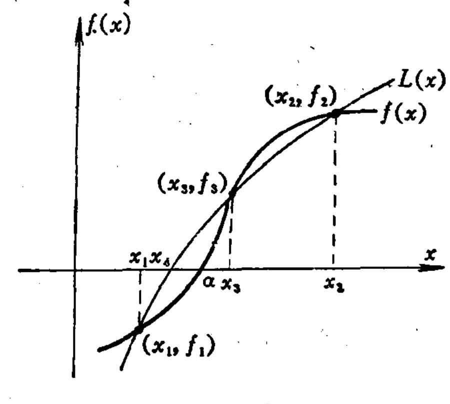


图 9.3

利用拉格朗日插值公式，作出通过(9.5.1)中三个点的二次函数

$$L(z)=\frac{(z-z_2)(z-z_3)}{(z_1-z_2)(z_1-z_3)}f_1+\frac{(z-z_1)(z-z_3)}{(z_2-z_1)(z_2-z_3)}f_2+\frac{(z-z_1)(z-z_2)}{(z_3-z_1)(z_3-z_2)}f_3 \tag{9.5.2}$$

为计算上的方便，引入新的变量

$$\lambda=\frac{z-z_3}{z_3-z_2} \tag{9.5.3}$$

且令

$$\left.\begin{aligned}\lambda_3&=\frac{z_3-z_2}{z_2-z_1}\\ \delta_3&=1+\lambda_3\end{aligned}\right\} \tag{9.5.4}$$

改写(9.5.2)为 λ 的二次式

$$L(\lambda)=a\lambda^2+b\lambda+c \tag{9.5.5}$$

式中

$$\left.\begin{aligned}a&=f_1\lambda_3^2-f_2\lambda_3\delta_3+f_3\lambda_3\\ b&=f_1\lambda_3^2-f_2\delta_3^2+f_3(\lambda_3+\delta_3)\\ c&=f_3\delta_3\end{aligned}\right\} \tag{9.5.6}$$

二次函数 $L(\lambda)$ 存在两个零点

$$\lambda=\frac{-2c}{b\pm\sqrt{b^2-4ac}} \tag{9.5.7}$$

取其模小的一个(亦即分母中的正负号应取使分母的模值为大的那个)，记为 λ_4。由 λ 的定义得

$$z_4=z_3+\lambda_4(z_3-z_2) \tag{9.5.8}$$

作为解的一个新的近似值。然后以 z_2、z_3、z_4 代替 z_1、z_2、z_3 重复上过程得到 z_5。如此等等，直到近似值充分接近于解案时为止。

9.5.2 方法的收敛性

如果 $f(z)$ 在解 α 附近存在连续的三阶导数，并且初始近似值充分接近于所求的解 α，那末二次插值法的迭代过程是收敛的，且其收敛速度，在解为单根且当 z_k 充分接近于解 α 时有：

$$|z_{k+1}-\alpha|\approx K|z_k-\alpha|^{1.84} \tag{9.5.9}$$

式中

$$K=\left|\left(-\frac{f'''(\alpha)}{6f'(\alpha)}\right)^{0.42}\right|$$

$f'(\alpha)$、$f'''(\alpha)$ 分别是 $f(z)$ 在解 α 处的一阶和三阶导数值。式(9.5.9)表明，在接近收敛的迭代进程中，近似值的误差将按阶为 1.84 的速度下降；在解为二重根且当 z_k 充分接近于解 α 时有：

$$|z_{k+1}-\alpha|\approx K|z_k-\alpha|^{1.23} \tag{9.5.10}$$

式中

$$K=\left|\left(-\frac{f'''(\alpha)}{3f''(\alpha)}\right)^{0.23}\right|$$

式(9.5.10)表明在二重根的情况下，二次插值法仍有超线性的收敛速度。收敛性的证明可详见资料[2]。

上述结论从理论上说明了当初始近似值充分接近解案时迭代过程的收敛性。从实际使用表明，二次插值法对初值的要求并不苛刻，即使初值不太好也常常能收敛。但在具体使用算法时，为保证方法的有效性，尚需注意以下细节处理。

9.5.3 方法的若干细节处理

(1) 方法的异常情况和处理

在迭代计算过程中,可能遇到使公式(9.5.7)的计算发生困难的异常情况。此时,须加特殊处理方能使迭代继续进行。在二次插值法中,共存在以下两种异常情况:

(i) $f_1=f_2=f_3$, 在几何上就是三点位于一水平直线上,此时作出的抛物线退化为平行于 x 轴的直线。此种情形在计算中将出现 $\lambda_4=\frac{-2c}{0}=\infty\ (c\neq0)$。

(ii) $z_3=z_1$, 即三点中有二点相重,此时出现 $\lambda_4=\frac{0}{0}$, 量值不定。

异常情况的处理办法是可任取 λ_4 值,譬如可取 $\lambda_4=1$, 继续进行迭代。

(2) 改善收敛性的处理

为避免迭代计算中发生超射,即新的近似值 z_4 离上一次的近似值 z_3 不适当地远,从而可能引起发散。我们在得到新近似值后,用条件

$$\left|\frac{f(z_4)}{f(z_3)}\right|\leqslant M \tag{9.5.11}$$

加以控制。此处 M 是一常数,可取 10 或其他定值。当条件不满足时,就缩小 λ_4, 以 $\frac{1}{2}\lambda_4$ 代替 λ_4, 并重复判断条件(9.5.11),直到条件满足为止。

(3) 上下溢处理

在求零点的过程中,函数 $f(z)$ 的量值变化是比较大的,特别是当 $f(z)$ 含有指数函数或含有 z 的高幂次时变化尤甚。因此,在用计算机实现算法时,常需考虑上下溢处理,以避免由此造成计算中断,或带来有效数字的损失而影响方法的有效性。处理办法无非是在作乘除运算前考虑引入比例因子。

9.5.4 计算步骤

(1) 准备:选定三个初始近似值 z_1、z_2、z_3, 计算相应的 $f(z)$ 值 f_1、f_2、f_3, 并计算

$$\lambda_3=\frac{z_3-z_2}{z_2-z_1}$$

(2) 迭代:计算

$$\begin{aligned}
\delta_3&=1+\lambda_3\\
a&=f_1\lambda_3^2-f_2\lambda_3\delta_3+f_3\lambda_3\\
b&=f_1\lambda_3^2-f_2\delta_3^2+f_3(\lambda_3+\delta_3)\\
c&=f_3\delta_3\\
\lambda_4&=\frac{-2c}{b\pm\sqrt{b^2-4ac}}
\end{aligned}$$

上式分母中的"±"号,是取分母的模值为大的一个,如遇异常情形,则取 $\lambda_4=1$。

于是得新近似值

$$z_4=z_3+\lambda_4(z_3-z_2)$$

计算 $f_4=f(z_4)$。如果 f_4 不满足 $|f_4|/|f_3|\leqslant M$(M 是取定的常数),则缩小 λ_4, 直到条件满足为止。

(3) 控制：如果 z_4 满足

$$|\delta|\leqslant\varepsilon \quad 或 \quad |f_4|\sim 0 \quad (\delta,\ \varepsilon 的含义见式(9.3.8))$$

则认为过程收敛，终止迭代，以 z_4 作为所求的解案，否则执行下步。

(4) 迭代准备：如果迭代次数超过某个上界，则认为过程不收敛，终止迭代，计算失败。否则，以 z_2、z_3、z_4、f_2、f_3、f_4、λ_4 分别代替 z_1、z_2、z_3、f_1、f_2、f_3、λ_3，转步(2)继续迭代。

§9.6 线性分式插值法(双曲插值法)

9.6.1 方法简述

设已知非线性方程

$$f(z)=0$$

解的三个近似值：z_1、z_2、z_3，设相应的 $f(z)$ 的函数值是：f_1、f_2、f_3。通过这三个近似值点

$$(z_1,\ f_1),\ (z_2,\ f_2),\ (z_3,\ f_3)$$

构造线性分式函数

$$W(z)=\frac{z-a}{bz+c} \tag{9.6.1}$$

经整理后可写成

$$bwz+cw-z+a=0 \tag{9.6.2}$$

其中，系数 a、b、c 满足下列线性方程组

$$\left.\begin{aligned} a+f_1z_1b+f_1c=z_1 \\ a+f_2z_2b+f_2c=z_2 \\ a+f_3z_3b+f_3c=z_3 \end{aligned}\right\} \tag{9.6.3}$$

(9.6.2)式在实数情形时代表一条双曲线，见图 9.4，故有双曲插值法之称。

图 9.4

在近似值附近，就以此线性分式函数 $W(z)$ 来近似非线性函数 $f(z)$，且以此线性分式的零点

$$z_4=z_3+\frac{(z_3-z_2)(z_3-z_1)f_3(f_2-f_1)}{(z_3-z_2)(f_1-f_3)f_2+(z_3-z_1)(f_3-f_2)f_1} \tag{9.6.4}$$

作为非线性方程解的一个新近似值。然后以 z_2、z_3、z_4 代替 z_1、z_2、z_3，重复上过程得到 z_5，如此等等，直到近似值充分接近解案时为止。

因(9.6.4)式的计算格式在接近收敛时包含了两个相近数之差，有可能影响数值计算的准确度，为此引入量

$$\lambda_3=\frac{z_3-z_2}{z_2-z_1},\quad \delta_3=1+\lambda_3 \tag{9.6.5}$$

且将迭代格式(9.6.4)改写为

$$z_4=z_3+\lambda_4(z_3-z_2) \tag{9.6.6}$$

其中

$$\lambda_4=\frac{f_3(\delta_3f_2-\lambda_3f_1)-f_3f_1}{f_3(\delta_3f_1-\lambda_3f_2)-f_2f_1} \tag{9.6.7}$$

公式(9.6.5)、(9.6.6)、(9.6.7)即是线性分式插值法的迭代公式。

9.6.2 方法的收敛性

如果$f(z)$在解α的邻近解析，并且初始近似值充分接近于所求的解，那末线性分式插值法的迭代过程是收敛的，且其收敛速度，在解为单根且当z_k充分接近于α时有

$$|z_{k+1}-\alpha| \approx K|z_k-\alpha|^{1.84} \tag{9.6.8}$$

式中

$$K=\left|\frac{1}{4}\left(\frac{f''(\alpha)}{f'(\alpha)}\right)^2-\frac{1}{6}\,\frac{f'''(\alpha)}{f'(\alpha)}\right|^{0.42}$$

$f'(\alpha)$、$f''(\alpha)$、$f'''(\alpha)$分别是$f(z)$在解α处的一阶、二阶、三阶导数值。式(9.6.8)表明在接近收敛的迭代过程中，近似值的误差将按阶为1.84的速度下降。但当解为多重时，收敛速度就降为线性的了。具体证明见[3]。

9.6.3 方法的异常情况和处理

在按公式(9.6.5)、(9.6.6)、(9.6.7)进行迭代计算中可能遇到以下数种异常情况：

(1) $f_3=f_2=f_1$，即三点位于一水平直线上，此时$\lambda_4=\dfrac{0}{0}$。

(2) $f_3=f_2$或$f_3=f_1$或$f_2=f_1$，即二点位于一水平直线上，此时双曲线(9.6.2)退化为二条平行于轴的相互垂直的直线，计算时将产生如下情况：

在$f_2=f_1$时，有$\lambda_4=0$，$z_4=z_3$，产生假收敛；

在$f_3=f_1$时，有$\lambda_4=-1$，$z_4=z_2$，下步计算中将出现异常情况(3)；

在$f_3=f_2$时，有$\lambda_4=-\left(1+\dfrac{1}{\lambda_3}\right)$，$z_4=z_1$，三点不变，并在下一步产生假收敛。

(3) $z_3=z_1$，即两点相重，此时$\lambda_4=\dfrac{0}{0}$。

从以上分析看出，当遇到这些异常情况时，如果不及时处理，就会发生计算困难或者产生假收敛等情况而使方法失败。简便的处理方法是任取λ_4值，譬如可取$\lambda_4=1$，继续进行迭代。

9.6.4 计算步骤

(1) 准备：选定三个初始近似值z_1、z_2、z_3，计算相应的$f(z)$值f_1、f_2、f_3，并计算：

$$\lambda_3=\frac{z_3-z_2}{z_2-z_1}$$

(2) 迭代：计算

$$\delta_3=1+\lambda_3$$

$$\lambda_4=\begin{cases}\dfrac{f_3(\delta_3 f_2-\lambda_3 f_1)-f_3 f_1}{f_3(\delta_3 f_1-\lambda_3 f_2)-f_2 f_1} & \text{非异常情况时}\\ 1 & \text{异常情况时}\end{cases}$$

得新近似值

$$z_4=z_3+\lambda_4(z_3-z_2)$$

计算

$$f_4=f(z_4)$$

(3) 控制：如果z_4满足

$$|\delta|\leqslant\varepsilon \text{ 或 } |f_4|\approx 0 \quad (\delta,\ \varepsilon \text{ 的含义见式(10.2.8)})$$

则认为过程收敛，终止迭代，以 z_4 作为所求的解案。否则执行步(4)。

(4) 迭代准备：如果迭代次数超过某个上限，则认为过程不收敛，终止迭代，计算失败。否则，以 z_2、z_3、z_4、f_2、f_3、f_4、λ_4 分别代替 z_1、z_2、z_3、f_1、f_2、f_3、λ_3 转步 2，继续迭代。

§9.7 求非线性方程全部解的处理方法

上面介绍了五种求函数零点的方法。这些方法每次都只求出一个零点，而在实际计算问题中，常常需要计算一函数在复平面上的全部零点，或者在某个域内的所有零点。因此，如何应用这些方法来完成问题的计算，将在本节中提供一、二种处理方法。

常见的问题一般有如下三类：

(1) 要求确定函数 $f(z)$ 在复平面上的全部零点，而其零点数目为已知，且为有限个。例如要确定一多项式的全部零点时即是此种情形；

(2) 要求确定函数 $f(x)$ 在区间 $[a, b]$ 上的全部实零点；

(3) 要求确定出函数 $f(z)$ 在复平面某个区域内的全部零点，而其零点数目事先并不知道。

对于最后一类问题是比较困难的，处理办法往往需根据具体问题确定，此处就不予详述了。对于前面两类问题，我们将分别结合二次插值法和线性分式插值法进行介绍。在应用其他方法时，其处理方法亦是相仿佛的。

9.7.1 应用二次插值法求函数 $f(z)$ 在复平面上的有限个零点

设已知 $f(z)$ 在复平面上有 n 个零点，可按照下列步骤应用二次插值法逐个地定出它们的量值。

(1) 首次初始值的确定

估计模为最小的零点的粗略位置，假定是 z_0，则取

$$z_0(1-k),\ z_0(1+k),\ z_0$$

作为三个初始近似值，此处 k 是取定的常数，譬如可取 $k=\frac{1}{4}$。如果 $z_0=0$，则取 $-k$, k, 0 作为三个初始近似值。

(2) 定解

按取定的初值，应用二次插值法确定零点的精确值。

(3) 求下一零点的过渡

在求出一零点 z_1 后，取

$$y(1-k),\ y(1+k),\ y$$

作为求下一零点的初始值，以期求出在 z_1 附近的另一零点，此处 $y=z_1(1+\Delta)$，其中 Δ 是取定的常数，例如可取 $\Delta=2^{-8}$。

为了不使零点求重(重复求出已求得的零点)，在求出一零点后，便将 $f(z)$ 降阶，将该零点从 $f(z)$ 中除去。

对于 $f(z)$ 为多项式的情形，可以应用综合除法执行降阶运算，其算法如下：

设已求出多项式

$$f(z)=z^n+a_1z^{n-1}+a_2z^{n-2}+\cdots+a_{n-1}z+a_n$$

的一个零点 z_1，现将此零点从 $f(z)$ 中除去，即除以因子 $(z-z_1)$。若记降阶后的多项式为

$$f_1(z)=\frac{f(z)}{(z-z_1)}\equiv z^{n-1}+b_1z^{n-2}+b_2z^{n-3}+\cdots+b_{n-2}z+b_{n-1}$$

则 $f(z)$ 的系数 $b_1, b_2, \cdots, b_{n-1}$ 可由下递推公式逐个定出

$$b_1=a_1+z_1$$

$$b_k=a_k+b_{k-1}\cdot z_1 \quad (k=2, 3, \cdots, n-1)$$

这样，在每求出一零点后，多项式的次数便降低一次，减少了计算函数值的工作量，但在降阶运算过程中带来了误差积累。如果求根的次序按从大到小的次序进行，则降阶过程中引入的误差对后面一些小根精度的影响将可能是严重的(参见[4])。但如果按从小到大的次序进行，则经误差分析表明此过程是稳定的，一般不会影响后求根的精确度。

对于非多项式的函数，一般不能明显执行降阶运算。而是在已求出一零点 z_1 后，用函数

$$f_1(z)=\frac{f(z)}{(z-z_1)}$$

来代替 $f(z)$。一般地，在已求出 r 个零点 $z_1, z_2, \cdots, z_r$ 后，用函数

$$f_r(z)=\frac{f_{r-1}(z)}{(z-z_r)}=\frac{f(z)}{(z-z_1)\cdots(z-z_r)}$$

来代替函数 $f(z)$。显然，函数 $f_r(z)$ 包含了 $f(z)$ 的除 $z_1, \cdots, z_r$ 外的所有零点，且仅包含这些零点。因此，只要 $z_1, \cdots, z_r$ 求得足够精确，一般就不会发生求重的现象。

重复执行步骤(2)、(3)，直到 n 个零点全部求出为止。

9.7.2 应用线性分式插值法求 $f(x)$ 在给定区间 $[a, b]$ 上的全部实零点

设 $f(x)$ 在 $[a, b]$ 上的零点均是单的，则可用下述过程定出 $f(x)$ 在 $[a, b]$ 上的全部零点。

在 $[a, b]$ 上适当选取一串分点

$$a=x_1<x_2<\cdots<x_n<x_{n+1}=b$$

将 $[a, b]$ 分割成 n 个小区间

$$[x_1, x_2], [x_2, x_3], \cdots, [x_n, x_{n+1}]$$

这些小区间足够地小，使得在每个小区间上仅可能包含 $f(x)$ 的一个零点。

逐个地计算 $f(x)$ 在这些小区间端点 $x_i, x_{i+1} (i=1, 2, \cdots, n)$ 上的函数值 f_i, f_{i+1}，并检验它们是否为零或相互异号，

若 $f(x_i)=0$，则 x_i 即为一零点；

若 f_i 与 f_{i+1} 异号，则 x_i 与 x_{i+1} 之间存在一零点，此时以 $x_i, x_{i+1}, \frac{x_i+x_{i+1}}{2}$ 作为初始近似值，用双曲插值法进行迭代，求出该零点的精确值。

对所有的小区间均按此执行，即可定出 $f(x)$ 在 $[a, b]$ 上的所有零点。

§9.8 方法的选择

针对具体的计算问题，如何选择一种适宜的方法，使得既有效而又能较快地完成计算工作。自然，选择的原则不外乎是：

(1) 方法是有效的，就是说，针对所解问题的性质和特点，选用一种能成功地得到解的方法。对于求零点的迭代解法来说，也就是针对所具备的函数性态和初值情况，选用一种能收敛到解的方法。

(2) 方法的工作量是较少的，确切地说，就是在一定的初值条件下，将零点确定到一定的精确度所花费的工作量是较少的。自然，这不单要考虑方法的收敛速度，而且也要考虑到每迭代一次所花费的工作量。因为前者标志每迭代一次所能获得的解的近似度的增长速度，后者反映为获得这一近似度的增长所花费的代价。再者，对于一高速收敛的方法，如果它迭代一次的工作量比另一种收敛速度较低的方法要多，那末，究竟哪种方法的效率高，总的工作量较少呢？为此，引入一种衡量标准，即所谓有效指数的概念。

在求解零点的过程中，其主要工作量是在函数或其导数的计算上。如果将一个函数或其导数的计算量作为一个计算单位，并假设某种方法每次迭代需花费m个计算单位，而具有r阶的收敛速度。方法的r阶收敛速度是以每次迭代花费m个计算单位的代价所获得的，因此平均一个计算单位所获得的收敛速度的阶将是

$$R=r^{\frac{1}{m}} \tag{9.8.1}$$

数R就称为方法的有效指数。

有效指数R是反映每花费一个计算单位所能获得的解的近似度的增长速度。例如牛顿迭代法为二阶收敛速度，$r=2$，每次迭代需要二个计算单位，$m=2$，故$R=2^{\frac{1}{2}}=1.414$。这表示牛顿迭代法，当近似值充分接近于解时，每花费一个计算单位，近似值的误差将按阶为1.414的速度下降。

(3) 方法简单，易于编制程序，减轻计算前的准备工作。

关于方法的选择，通常需针对具体问题的特点和要求，根据上述原则权衡选择之。表9.2对上面介绍的几种方法作出了比较：

表 9.2

方法名称	收敛速度r（单根时）	有效指数R	重根时的收敛速度
区间分半法	1	1	1，偶重时失败
线性插值法	1.618	1.618	1
牛顿法	2	1.414	1
二次插值法	1.84	1.84	二重时为1.23
线性分式插值法	1.84	1.84	1

由表9.2看出，牛顿法的收敛速度最高，但从有效指数看，其效率不如二次插值法和线性分式插值法。且因为牛顿法用到了一阶导数$f'(x)$，故它对易写出导数式的函数适宜，特别是当同时计算函数及其导数值并不比单独计算函数值增加太多工作量时，牛顿法的效率就比较高。在收敛性方面，从实际使用表明，它对初始值的要求比较苛刻。因此，必须具备良好的初值。

区间分半法是一种低效率的方法，且只能求实根。其优点是方法简单，且对函数$f(x)$的要求比较低，仅需$f(x)$本身连续。因此适用于光滑程度差的函数。

上列方法中，效率最高的是二次插值法和线性分式插值法。它们具有相同的收敛速度和有效指数。但遇二重根时，二次插值法还保持了超线性的收敛速度，优于线性分式插值法。二次插值法的最大优点是：从实际使用表明，它对初值的要求不苛刻，即使用比较坏的初值，亦常可获得收敛。它的缺点是编制程序较为复杂，并且即使计算实零点亦常需采用复运算过程，增加了不必要的工作量。因此二次插值法用于求复零点较宜，特别是在无良好初值的情形。线性分式插值法，在求实零点时其运算过程可以全部采用实运算。在收敛性方面，从实际使用看，求实零点较求复零点要好些。因此线性分式插值法用于求实零点较好。

§9.9 非线性方程组的解法

考虑非线性方程组

$$\left.\begin{aligned} f_1(x_1,\ x_2,\ \cdots,\ x_n)=0 \\ f_2(x_1,\ x_2,\ \cdots,\ x_n)=0 \\ \vdots \\ f_n(x_1,\ x_2,\ \cdots,\ x_n)=0 \end{aligned}\right\} \tag{9.9.1}$$

此处 $x_1,\ x_2,\ \cdots,\ x_n$ 是实变量，$f_i(i=1,\ 2,\ \cdots,\ n)$是未知量 $x_1,\ x_2,\ \cdots,\ x_n$ 的非线性实函数。要求确定方程组(9.9.1)在指定范围内的一组解 $\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_n$。

解决此类问题的方法，常用的有如下二种：一种是属于线性化的方法，即是将非线性方程组以一线性方程组来近似，由此构造一种迭代格式，用以逐次逼近所求的解案；另一种是属于求函数极小值的方法，即是由这些非线性函数 $f_1,\ f_2,\ \cdots,\ f_n$ 构造一模函数 Φ，例如可构造模函数 $\Phi(x_1,\ \cdots,\ x_n)=\sum\limits_{i=1}^{n}[f_i(x_1,\ \cdots,\ x_n)]^2$。然后，以各种各样的下降法求出模函数的极小值点，而此极小值点即是非线性方程组的一组解。

下面介绍三种解法，一是牛顿迭代法，它是属于线性化的方法，其他二种是最速下降法和变尺度法中的 DFP 法，它们是属于求极小值点的方法。

为了叙述清楚起见，现以一个二阶方程组为例来介绍这些方法。对于一般的情形，方法是不难类推的。下面考虑非线性方程组

$$\left.\begin{aligned} f_1(x,\ y)=0 \\ f_2(x,\ y)=0 \end{aligned}\right\} \tag{9.9.2}$$

此处，$x,\ y$ 是实变量；$f_1,\ f_2$ 是未知量 $x,\ y$ 的非线性实函数。同时，假定f_1、f_2 对 x、y 的二阶偏导数存在且连续，并且在解的邻近，行列式

$$\boldsymbol{J}=\det\begin{pmatrix} \dfrac{\partial f_1}{\partial x}, & \dfrac{\partial f_1}{\partial y} \\ \dfrac{\partial f_2}{\partial x}, & \dfrac{\partial f_2}{\partial y} \end{pmatrix}\neq 0 \tag{9.9.3}$$

§9.10 解非线性方程组的牛顿迭代法

设已知方程组(9.9.2)之解的一组初始近似值 $x_0,\ y_0$。在此近似值邻近将非线性函数 f_1、f_2 用如下的线性函数来近似：

$$\left.\begin{aligned} f_1(x,\ y) &\approx f_1(x_0,\ y_0)+(x-x_0)\frac{\partial f_1}{\partial x}(x_0,\ y_0)+(y-y_0)\frac{\partial f_1}{\partial y}(x_0,\ y_0) \\ f_2(x,\ y) &\approx f_2(x_0,\ y_0)+(x-x_0)\frac{\partial f_2}{\partial x}(x_0,\ y_0)+(y-y_0)\frac{\partial f_2}{\partial y}(x_0,\ y_0) \end{aligned}\right\} \tag{9.10.1}$$

其中，$\frac{\partial f_1}{\partial x}(x_0,\ y_0)$表示$f_1$对$x$的一阶偏导数$\frac{\partial f_1}{\partial x}$在$x_0$、$y_0$处的取值，其他类似符号的意义亦相仿。实际上，近似式(9.10.1)是由f_1、f_2在$(x_0,\ y_0)$处的泰劳展式略去其高于线性的项后所得的。

于是得到一线性方程组

$$\left.\begin{aligned} \frac{\partial f_1}{\partial x}(x_0,\ y_0)\Delta x+\frac{\partial f_1}{\partial y}(x_0,\ y_0)\Delta y &= -f_1(x_0,\ y_0) \\ \frac{\partial f_2}{\partial x}(x_0,\ y_0)\Delta x+\frac{\partial f_2}{\partial y}(x_0,\ y_0)\Delta y &= -f_2(x_0,\ y_0) \end{aligned}\right\} \tag{9.10.2}$$

式中
$$\Delta x=x-x_0,\ \Delta y=y-y_0$$

由前面的假定，行列式

$$\boldsymbol{J}=\det\begin{pmatrix} \frac{\partial f_1}{\partial x}, & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x}, & \frac{\partial f_2}{\partial y} \end{pmatrix}$$

在解的邻近不为零，因此只要近似值$(x_0,\ y_0)$充分靠近于解，那末方程组(9.10.2)的系数行列式是不为零的，于是可以从(9.10.2)中解出Δx，Δy。由它们的含义得到

$$\left.\begin{aligned} x_1 &= x_0+\Delta x \\ y_1 &= y_0+\Delta y \end{aligned}\right\} \tag{9.10.3}$$

以此作为非线性方程组解的一个新近似值。然后，以x_1，y_1代替x_0，y_0，重复上述过程，直到相邻二次近似值x_k，y_k和x_{k+1}，y_{k+1}满足条件：

$$\max\{\delta_x,\ \delta_y\}<\varepsilon \tag{9.10.4}$$

其中

$$\delta_x=\begin{cases} \frac{|x_{k+1}-x_k|}{|x_k|} & \text{当 } |x_k|\geqslant c \\ |x_{k+1}-x_k| & \text{当 } |x_k|<c \end{cases}$$

$$\delta_y=\begin{cases} \frac{|y_{k+1}-y_k|}{|y_k|} & \text{当 } |y_k|\geqslant c \\ |y_{k+1}-y_k| & \text{当 } |y_k|<c \end{cases}$$

或者满足条件：

$$\max\{|f_1|,\ |f_2|\}<\delta \tag{9.10.5}$$

时为止。最后得到的近似值即作为所要的解案。此处，ε是允许误差，c是取绝对或相对误差的控制数，它们随具体问题的要求而定；δ是一接近于零的小数，视计算机的字长和数值范围而定。

在方程组(9.9.3)的假设条件下，可以证明当初值充分接近于解时，牛顿迭代过程是按平方收敛速度收敛的。但在实际使用中表明，牛顿迭代法需要有较好的初值，否则很有可能发散。

§9.11 最速下降法

对于方程组(9.9.2)，构造模函数

$$\Phi=\Phi(x, y)=[f_1(x, y)]^2+[f_2(x, y)]^2 \tag{9.11.1}$$

显然方程组(9.9.2)之解是 Φ 的零极小值点，反之亦然。因此可通过求 Φ 的零极小值点来得到方程组(9.9.2)之解。

函数 $\Phi(x, y)$ 在几何上是一空间曲面，它与 x-y 面相切的点即是它的零极小值点(见图 9.5)。

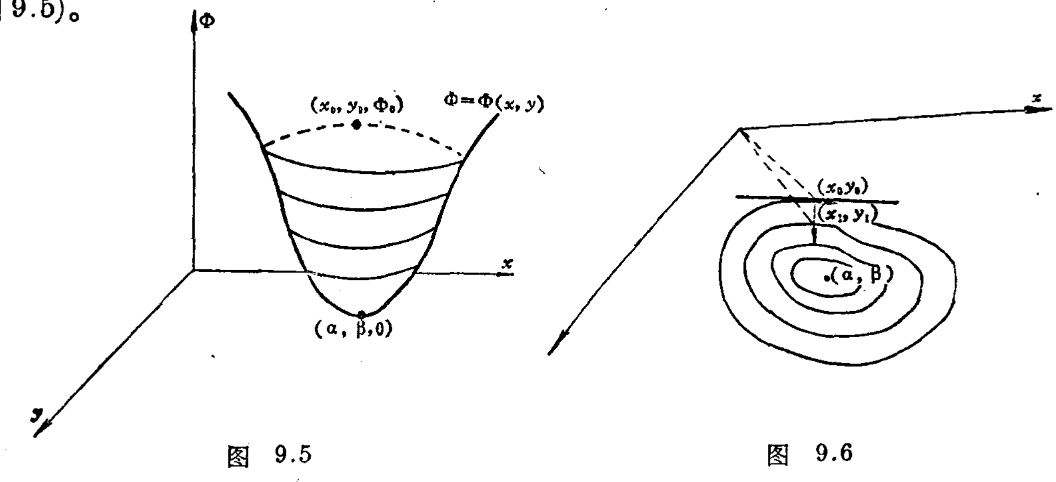


图 9.5　　　　图 9.6

对于空间曲面 $\Phi=\Phi(x, y)$，如果用一系列平行于 x-y 面的平面 $\Phi=$常数相截之，可以得到一族平面曲线，将它们投影于 x-y 面，得到如图 9.6 所示的曲线族，称为曲面的等高线族，处在同一条等高线上的点其 Φ 值是相同的。每条等高线相应于一个 Φ 值，在极小值点 (α, β) 附近，其等高线形成以 (α, β) 为中心的封闭曲线族，且相应的 Φ 值由外往里不断地下降，当达到 (α, β) 时，Φ 值为零。

对于处在函数 $\Phi(x, y)$ 定义域 D 内的任何点，总存在族中的一条等高线通过它。如果从 D 内的某个点 (x_0, y_0) 出发沿着使 Φ 值下降的方向逐步地下降 Φ 值，一直降到它的零极小值，那末就可以得到所求问题的解。

我们知道，在一点处等高线的法向，也就是函数 $\Phi(x, y)$ 在该点处的梯度方向

$$\boldsymbol{g}=\left(\frac{\partial\Phi}{\partial x}, \frac{\partial\Phi}{\partial y}\right)^T$$

是使 Φ 值上升最快的方向，因此其反方向

$$-\boldsymbol{g}=\left(-\frac{\partial\Phi}{\partial x}, -\frac{\partial\Phi}{\partial y}\right)^T$$

就是使 Φ 值下降最快的方向。最速下降法就是沿着这样的方向来逐步下降 Φ 值的。具体方法如下：

设 (x_0, y_0) 是解的一个近似值。计算 Φ 在此点的梯度

$$\boldsymbol{g}_0=(g_{10}, g_{20})^T$$

此处

$$\left.\begin{aligned} g_{10}&=\left(\frac{\partial\Phi}{\partial x}\right)_0=2\left[\left(\frac{\partial f_1}{\partial x}\right)_0(f_1)_0+\left(\frac{\partial f_2}{\partial x}\right)_0(f_2)_0\right] \\ g_{20}&=\left(\frac{\partial\Phi}{\partial y}\right)_0=2\left[\left(\frac{\partial f_1}{\partial y}\right)_0(f_1)_0+\left(\frac{\partial f_2}{\partial y}\right)_0(f_2)_0\right] \end{aligned}\right\} \tag{9.11.2}$$

其中下标"0"表示括号内的函数在(x_0, y_0)处的取值。

从这点出发沿负梯度方向，$-\boldsymbol{g}_0$，跨一适当的步长，得到新的点

$$\left.\begin{aligned} x_1 &= x_0 - \lambda g_{10} \\ y_1 &= y_0 - \lambda g_{20} \end{aligned}\right\} \tag{9.11.3}$$

此处因子λ作如此选择，使得新的点x_1，y_1是$\Phi(x, y)$在此方向$-\boldsymbol{g}_0$上的相对极小值点，即有

$$\Phi(x_1, y_1) = \min_{\lambda}\{\Phi(x_0 - \lambda g_{10}, y_0 - \lambda g_{20})\}$$

因为目的是要定出在x_0, y_0附近的另一近似值，故可以将$f_i(x_0 - \lambda g_{10}, y_0 - \lambda g_{20})$ $(i = 1, 2)$在$\lambda = 0$处展开并略去λ^2及以后的项，得函数Φ的近似式

$$\begin{aligned} \Phi(x_0 - \lambda g_{10}, y_0 - \lambda g_{20}) \approx [(f_1)_0^2 + (f_2)_0^2] &- 2\lambda\left\{\left[\left(\frac{\partial f_1}{\partial x}\right)_0 g_{10} + \left(\frac{\partial f_1}{\partial y}\right)_0 g_{20}\right](f_1)_0\right. \\ &\left. + \left[\left(\frac{\partial f_2}{\partial x}\right)_0 g_{10} + \left(\frac{\partial f_2}{\partial y}\right)_0 g_{20}\right](f_2)_0\right\} \\ &+ \lambda^2\left\{\left[\left(\frac{\partial f_1}{\partial x}\right)_0 g_{10} + \left(\frac{\partial f_1}{\partial y}\right)_0 g_{20}\right]^2\right. \\ &\left. + \left[\left(\frac{\partial f_2}{\partial x}\right)_0 g_{10} + \left(\frac{\partial f_2}{\partial y}\right)_0 g_{20}\right]^2\right\} \end{aligned}$$

若记

$$\boldsymbol{J}_0 = \begin{pmatrix} \left(\frac{\partial f_1}{\partial x}\right)_0, & \left(\frac{\partial f_1}{\partial y}\right)_0 \\ \left(\frac{\partial f_2}{\partial x}\right)_0, & \left(\frac{\partial f_2}{\partial y}\right)_0 \end{pmatrix} \qquad \boldsymbol{F}_0 = \begin{pmatrix} (f_1)_0 \\ (f_2)_0 \end{pmatrix} \qquad \boldsymbol{g}_0 = \begin{pmatrix} g_{10} \\ g_{20} \end{pmatrix} = 2J_0^T\boldsymbol{F}_0$$

上标T表示转置，且记(a, b)为向量a, b的内积，则有

$$\Phi(x_0 - \lambda g_{10}, y_0 - \lambda g_{20}) \approx (\boldsymbol{F}_0, \boldsymbol{F}_0) - 2\lambda(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{F}_0) + \lambda^2(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{J}_0\boldsymbol{g}_0)$$

根据一元函数极小值的求法，解方程

$$\frac{\partial\Phi}{\partial\lambda} \approx -2(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{F}_0) + 2\lambda(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{J}_0\boldsymbol{g}_0) = 0$$

得

$$\lambda = \frac{(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{F}_0)}{(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{J}_0\boldsymbol{g}_0)} = \frac{(\boldsymbol{g}_0, \boldsymbol{g}_0)}{2(\boldsymbol{J}_0\boldsymbol{g}_0, \boldsymbol{J}_0\boldsymbol{g}_0)} \tag{9.11.4}$$

将它代入(9.11.3)得到x_1, y_1，以此作为Φ沿$-\boldsymbol{g}_0$方向的相对极小值点的近似值。如果由(9.11.4)得到的λ有

$$\Phi(x_1, y_1) < \Phi(x_0, y_0) \tag{9.11.5}$$

则就以它作为λ的取值，否则可缩小λ，例如可每次缩小一半，直到条件(9.11.5)满足为止。

将最后确定的λ值代入式(9.11.3)便得到新近似值x_1, y_1。然后再从新近似值(x_1, y_1)出发重复执行上过程。如此不断地进行，直到Φ值(在Φ值比较小时一般改用$\Psi = |f_1| + |f_2|$值，因在机器计算中当接近于解时，Φ值很快趋于零而消失了数字)降到充分小时为止，最后得到的近似值即作为所计算的解。

一般说来，最速下降法对任意初值都能收敛，但其收敛速度却是线性的。在开始几步后，其收敛速度就变得十分缓慢，尤其是在解案邻近，常常为了提高一点精度而需要付出很大的代价。因此，在实际使用上，常与牛顿法结合应用，使两个方法相互取长补短，以达到既

能保证收敛性又能加快收敛速度的目的。结合的方式是多种多样的。最原始的结合方式是：开始采用最速下降法，而当 Φ 值降到一定程度时，便改用牛顿法。

对于难于求出函数之偏导数或者偏导数的计算过于繁复的问题，可以用差商来近似微商。例如可用

$$\left.\begin{aligned}\frac{f_i(x+h,\ y)-f_i(x,\ y)}{h}\approx\frac{\partial f_i}{\partial x}\\ \frac{f_i(x,\ y+h)-f_i(x,\ y)}{h}\approx\frac{\partial f_i}{\partial y}\end{aligned}\right\}\quad (i=1,\ 2)$$

关于步长 h 的取法，有一种方案是：开始取 $h_0=10^{-3}$，在第 k 次迭代时取

$$h_k=\min\left\{10^{-3},\ \frac{1}{10}\max(|\Delta x_{k-1}|,\ |\Delta y_{k-1}|)\right\}$$

其中 $\Delta x_{k-1}=x_{k-1}-x_{k-2}$；$\Delta y_{k-1}=y_{k-1}-y_{k-2}$ 为上一次的迭代修正量。

§9.12 DFP 方 法

与最速下降法同样，将解非线性方程组

$$f_i(x_1,\ \cdots,\ x_n)=0\quad (i=1,\ 2,\ \cdots,\ n)\tag{9.12.1}$$

的问题化为定函数

$$\Phi(\boldsymbol{X})=\sum_{i=1}^{n}f_i(\boldsymbol{X})^2\tag{9.12.2}$$

的极小值的问题，此处 $\boldsymbol{X}=(x_1,\ \cdots,\ x_n)^T$ 是由未知量组成的向量，上标 T 表示转置。

$\Phi(\boldsymbol{X})$ 是 n 个未知量的非线性函数，如果在极小值点 $\boldsymbol{X}_{\min}$ 附近存在连续的三阶导数，则利用泰劳展式可知，在极小值点 $\boldsymbol{X}_{\min}$ 附近，$\Phi(\boldsymbol{X})$ 与一二次函数

$$\Phi(\boldsymbol{X})\approx\Phi(\boldsymbol{X}_0)+(\boldsymbol{g}_0,\ \boldsymbol{X}-\boldsymbol{X}_0)+\frac{1}{2}(\boldsymbol{X}-\boldsymbol{X}_0,\ \boldsymbol{H}_0(\boldsymbol{X}-\boldsymbol{X}_0))\tag{9.12.3}$$

相近似，此处，$\boldsymbol{X}_0$ 是在极小值附近的某个点；$\boldsymbol{g}_0$ 是 $\Phi(\boldsymbol{X})$ 在 $\boldsymbol{X}_0$ 处的梯度值：

$$\boldsymbol{g}_0=\left(\frac{\partial\Phi}{\partial x_1},\ \cdots,\ \frac{\partial\Phi}{\partial x_n}\right)^T_{\boldsymbol{X}=\boldsymbol{X}_0}$$

$$\boldsymbol{H}_0=\left(\frac{\partial\Phi}{\partial x_i\partial x_j}\bigg|_{\boldsymbol{X}=\boldsymbol{X}_0}\right)$$

是由 $\Phi(\boldsymbol{X})$ 在 $\boldsymbol{X}_0$ 处的二阶偏导数值所组成的对称矩阵

$$\boldsymbol{H}_0=\begin{pmatrix}\frac{\partial\Phi}{\partial x_1\partial x_1}, & \frac{\partial\Phi}{\partial x_1\partial x_2}, & \cdots, & \frac{\partial\Phi}{\partial x_1\partial x_n}\\ \frac{\partial\Phi}{\partial x_1\partial x_2}, & \frac{\partial\Phi}{\partial x_2\partial x_2}, & \cdots, & \frac{\partial\Phi}{\partial x_2\partial x_n}\\ \cdots\cdots & & & \\ \frac{\partial\Phi}{\partial x_1\partial x_n}, & \frac{\partial\Phi}{\partial x_2\partial x_n}, & \cdots, & \frac{\partial\Phi}{\partial x_n\partial x_n}\end{pmatrix}_{\boldsymbol{X}=\boldsymbol{X}_0}$$

在极小值附近它是正定的。

对于二次函数(9.12.3)，根据极值原理(在极值处的梯度为零)，由方程

$$\boldsymbol{g}(\boldsymbol{X})\equiv\boldsymbol{g}_0+\boldsymbol{H}_0(\boldsymbol{X}-\boldsymbol{X}_0)=0\tag{9.12.4}$$

可立即得出(9.12.3)之极小值的位置

$$\boldsymbol{X}=\boldsymbol{X}_0-\boldsymbol{H}_0^{-1}\boldsymbol{g}_0\tag{9.12.5}$$

因此，如果已知函数 $\Phi(\boldsymbol{X})$ 在极小值邻近一点处的梯度和二阶导数值，那末由(9.12.5)即可得到 $\boldsymbol{X}_{\min}$ 的更好近似值；并且，如果以此作为一种迭代格式，那末它将具有二次收敛速度。但遗憾的是，它在计算中用到了 Φ 的二阶导数值，这在实际计算中是十分麻烦的。因此，如果能避免二阶导数的计算而又能很快地定出一二次函数的极小值点，那末就能得到一种切实可行的快速收敛算法。求极小值的 DFP 方法就是这样的一种方法。其具体方法如下：

考虑二次函数

$$F(\boldsymbol{X})=a+(\boldsymbol{b},\ \boldsymbol{X})+\frac{1}{2}(\boldsymbol{X},\ \boldsymbol{H}\boldsymbol{X}) \tag{9.12.6}$$

式中，$\boldsymbol{X}$ 是由 n 个未知量组成的向量；a 是常数；$\boldsymbol{b}$ 是常向量；$\boldsymbol{H}$ 是对称正定的常矩阵。

易于写出 $F(\boldsymbol{X})$ 的梯度为

$$\boldsymbol{g}(\boldsymbol{X})=\boldsymbol{b}+\boldsymbol{H}\boldsymbol{X} \tag{9.12.7}$$

设 $\boldsymbol{X}_0$ 是任取的一个点，则由此点上的梯度值 $\boldsymbol{g}_0$ 即可得 $F(\boldsymbol{X})$ 的极小值的位置

$$\boldsymbol{X}_{\min}=\boldsymbol{X}_0-\boldsymbol{H}^{-1}\boldsymbol{g}_0 \tag{9.12.8}$$

为避免使用矩阵 $\boldsymbol{H}$ 和求逆计算，任取正定对称的矩阵 $\boldsymbol{G}_0$ 作为矩阵 $\boldsymbol{H}^{-1}$ 的初始试验值，并用公式

$$\boldsymbol{X}_{i+1}=\boldsymbol{X}_i+\alpha_i\boldsymbol{p}_i \tag{9.12.9}$$

$$\boldsymbol{p}_i=-\boldsymbol{G}_i\boldsymbol{g}_i \tag{9.12.10}$$

代替式(9.12.8)，作为求 $\boldsymbol{X}_{\min}$ 的迭代公式，此处，$\boldsymbol{X}_i$，$\boldsymbol{X}_{i+1}$ 分别是第 i 次和第 $i+1$ 次的近似值，而 $\boldsymbol{p}_i$ 是修正方向。其中，$\boldsymbol{g}_i=\boldsymbol{g}(\boldsymbol{X}_i)$ 是函数 $F(\boldsymbol{X})$ 在 $\boldsymbol{X}_i$ 处梯度；$\boldsymbol{G}_i$ 是第 i 次迭代所用的矩阵 $\boldsymbol{H}^{-1}$ 的试验值；α_i 是步长。

确定步长 α_i，使得 $\boldsymbol{X}_{i+1}$ 是 $F(\boldsymbol{X})$ 在直线

$$\boldsymbol{X}=\boldsymbol{X}_i+\lambda\boldsymbol{p}_i \tag{9.12.11}$$

上的极小值点。即使 α_i 满足条件

$$\frac{d}{d\lambda}F(\boldsymbol{X}_i+\lambda\boldsymbol{p}_i)\big|_{\lambda=\alpha_i}=0 \tag{9.12.12}$$

亦即满足

$$(\boldsymbol{g}_{i+1},\ \boldsymbol{p}_i)=0 \tag{9.12.13}$$

式(9.12.13)表明修正方向 $\boldsymbol{p}_i$ 与 $\boldsymbol{X}_{i+1}$ 处的梯度方向 $\boldsymbol{g}_{i+1}$ 相正交。关于求 α_i 的数值方法，将在后面介绍。

在 α_i 确定后，由式(9.12.9)即可得新近似值 $\boldsymbol{X}_{i+1}$。然后修正矩阵 $\boldsymbol{G}_i$，使修正后的矩阵

$$\boldsymbol{G}_{i+1}=\boldsymbol{G}_i+\Delta\boldsymbol{G}_i \tag{9.12.14}$$

满足条件：

(1) $\boldsymbol{G}_{i+1}$ 是正定对称的；

(2) $\boldsymbol{G}_{i+1}$ 满足

$$\boldsymbol{G}_{i+1}\boldsymbol{H}\boldsymbol{p}_j=\boldsymbol{p}_j\quad(j=0,\ 1,\ \cdots,\ i) \tag{9.12.15}$$

即所有前面的修正方向 $\boldsymbol{p}_0$，$\boldsymbol{p}_1$，$\cdots$，$\boldsymbol{p}_i$ 均是矩阵 $\boldsymbol{G}_{i+1}\boldsymbol{H}$ 的对应于单位特征值的特征向量。

条件(1)是使算法的修正方向 $\boldsymbol{p}_i=-\boldsymbol{G}_i\boldsymbol{g}_i$ 均是下降方向。这保证了算法的稳定性，即算法能稳定地逐步降低 $F(\boldsymbol{X})$ 值；条件(2)与式(9.12.13)保证了算法对于 n 元的二次函数，

将最多在 n 次迭代后达到极小值，并且有 $G_n=H^{-1}$。事实上可以证明，修正方向 p_0, p_1, p_2, …, p_{n-1} 是相互 H 共轭的，因而是线性无关的，且均正交于第 n 次近似值 X_n 处的梯度方向 g_n，即有关系式

$$\left.\begin{aligned}(p_i, Hp_j)=0 \quad & i, j=0, 1, \cdots, n-1 \quad i\neq j\\ (g_n, p_j)=0 \quad & j=0, 1, \cdots, n-1\end{aligned}\right\} \tag{9.12.16}$$

因此，根据线性代数中的以下结论：在 n 维空间中任一与 n 个线性无关的非零向量相正交的向量必是零向量。可以知道，必有 $g_n=0$，即 X_n 为极小值点。因此，对于二次函数来说，其算法仅以不超过未知量数目的步数，便可达到极小值。并且，因 p_1, …, p_n 是矩阵 HG_n 的 n 个具有单位特征值的线性无关的特征向量，故知 HG_n 必为单位矩阵，于是有 $G_n=H^{-1}$。

满足条件(1)、(2)的矩阵修正方法不是唯一的。在 Fletcher-Powell 的算法中是取如下的矩阵修正公式：

$$G_{i+1}=G_i+\frac{\Delta X_i\Delta X_i^T}{(\Delta X_i, \Delta g_i)}-\frac{G_i\Delta g_i(G_i\Delta g_i)^T}{(\Delta g_i, G_i\Delta g_i)} \tag{9.12.17}$$

此处，$\Delta X_i=X_{i+1}-X_i$; $\Delta g_i=g_{i+1}-g_i$; 上标 T 表示该向量的转置。可以证明，按此公式修正的矩阵是满足上述条件的，具体证明可见[6]。

公式(9.12.9)，(9.12.10)，(9.12.17)就是 DFP 方法的基本迭代公式。对于二次函数，它将以不超过变元数的迭代次数达到极小值。对于非二次函数，亦将具有快速的收敛速度。

以上就是方法的大致轮廓，对于实际使用还有如下两个问题：

(1) 初始矩阵 G_0 的选择。一般可取单位矩阵，这样，初始修正方向 p_0 将落在 X_0 处的最速下降方向 $-g_0$ 上。

(2) 步长 α_i 的实际计算法，就是定一元函数

$$y(\lambda)\equiv F(X_i+\lambda p_i)$$

之极小值的计算方法。一般是采用某种近似求法，即构造 $y(\lambda)$ 的近似多项式，然后求出其极小值点。此方面的处理方法是很多的，有一种方案是：在直线 $X=X_i+\lambda p_i$ 上取两个点：$\lambda=a$ 和 $\lambda=b$，使得在此两个点中存在 $y(\lambda)$ 的一个极小值，即有

$$y'(a)=(p_i, g(X_i+ap_i))<0$$
$$y'(b)=(p_i, g(X_i+bp_i))>0$$

然后由 a, b 处的函数值及导数值 $y(a)$、$y(b)$、$y'(a)$、$y'(b)$，利用埃尔米特插值多项式，得到一个三次多项式。求其导数后得

$$H_3'(\lambda)=\frac{(\lambda-b)^2}{(b-a)^2}[y'(a)+y'(b)+2z]+\frac{2(\lambda-b)}{b-a}[y'(b)+z]+y'(b)$$

其中

$$z=\frac{3[y(a)-y(b)]}{b-a}+y'(a)+y'(b) \tag{9.12.18}$$

易于定出 $H_3'(\lambda)$ 之相应于 $H(\lambda)$ 极小值的根，于是得 α_i 之近似值

$$\alpha_i\approx b-\frac{y'(b)+z-w}{y'(a)+y'(b)+2z}(b-a)$$

其中

$$w=\sqrt{z^2-y'(a)y'(b)} \tag{9.12.19}$$

但因 $y'(a)$ 与 $y'(b)$ 异号，且在极小值邻近又接近于零，为改善计算效果可改用算式

$$\alpha_i = b - \frac{y'(b) + w - z}{y'(b) - y'(a) + 2w}(b-a) \tag{9.12.20}$$

关于点 a、b 的选择，可按下述方案确定：取如下的一串 λ 值

$$\lambda = 0,\ h,\ 2h,\ 4h,\ 8h,\ \cdots$$

每个 λ 值是前一次的加倍，而

$$h = \min(k,\ (\boldsymbol{p}_i,\ \boldsymbol{p}_i)^{-\frac{1}{2}})$$

其中

$$k = -\frac{2(F(\boldsymbol{X}_i) - F(\boldsymbol{X}_{\min}))}{(\boldsymbol{g}_i,\ \boldsymbol{p}_i)}$$

k 值的意义是：对于二次函数，当 $\boldsymbol{G}_i = \boldsymbol{H}^{-1}$ 时，$k\boldsymbol{p}_i$ 将是 $\boldsymbol{X}_i$ 到极小值点 $\boldsymbol{X}_{\min}$ 的距离。

依次计算在这些 λ 上的 $y'(\lambda)$ 值。因 $y'(0) < 0$，故当遇某个 $\lambda = 2ih$ 有 $y'(2ih) > 0$ 时，即表明在 $\lambda = ih$ 与 $\lambda = 2ih$ 间存在 $y(\lambda)$ 的一个极小值。于是可取 $a = ih$，$b = 2ih$。

DFP 方法是最早的变尺度法，也是得到广泛使用的方法之一。但 DFP 算法存在以下的弱点：一是矩阵 $\boldsymbol{H}_i$ 可能出现奇异。Bard[7] 从分析修正公式(9.12.17)表明：当模 $\|\Delta\boldsymbol{x}_i\|$ 与 $\|\Delta\boldsymbol{g}_i\|$ 的量级相差悬殊时，矩阵 $\boldsymbol{H}_{i+1}$ 将出现奇异，并且一当 $\boldsymbol{H}_i$ 奇异，以后的矩阵将恒为奇异，于是矩阵的正定性不能保证，步长 α 出现负值甚至为零而产生假收敛；二是 Powell[8] 证明了 $(\boldsymbol{g}_i,\ \boldsymbol{H}_i\boldsymbol{g}_i)$ 是单调下降的。因此当某次 $\boldsymbol{x}_i$ 跨近鞍点时，因 $\boldsymbol{g}_i \sim 0$，有 $(\boldsymbol{g}_i,\ \boldsymbol{H}_i\boldsymbol{g}_i) \sim 0$，于是在以后的迭代中 $\|\boldsymbol{Hg}\| = \frac{1}{|\alpha|}\|\Delta\boldsymbol{x}\|$ 将被限制得更小，而可能陷入被吸引在鞍点邻近而无法摆脱的困境；三是方法的有效性与定一元函数极小即解方程(9.12.12)的精度有关，精度太低时常招致方法失败，但要精度高就要花费工作量。

为改善 DFP 算法的效果，常循环地让矩阵 $\boldsymbol{H}$ 重取初值。例如可每迭代 $n+1$ 次，舍去当前的矩阵 $\boldsymbol{H}_n$，改用初始矩阵 $\boldsymbol{H}_0(=I)$ 重新开始。

DFP 方法虽则存在上述问题，但从大量应用表明，在变尺度法中仍不失为一种有效的算法。

附录　解非线性方程和方程组程序

一、区间分半法程序

HITL(*a*, *b*, *eps*, *x*, *fx*, *FUNC*)

使用说明

过程 *HITL* 是用区间分半法求连续函数 $f(x)$ 在区间 $[a, b]$ 上的一个实零点。

输入参数:

a, b——定解区间的端点。要求 $f(a)$ 与 $f(b)$ 异号;

eps——允许误差。当逐次分半后的区间长度 $\Delta < eps$ 时，过程终止，以区间的中点作为所求的根值，此处

$$\Delta = \begin{cases} |b_k - a_k| & \text{当 } \frac{1}{2}|a_k + b_k| < 1 \\ |b_k - a_k| \Big/ \frac{1}{2}|a_k + b_k| & \text{当 } \frac{1}{2}|a_k + b_k| \geqslant 1 \end{cases}$$

FUNC——是计算函数值 $f(x)$ 的过程。需有过程说明: 过程 *FUNC*(x, f); 值 x; 简变 f; 始…终;

输出参数:

x——零点计算值;

fx——$fx = f(x)$。

程序

```
过程 HITL(a, b, eps, x, fx, FUNC);
  值 a, b, eps;
  简变 x, fx;
  过程 FUNC;
始  简变 FA, FB, SF, DX;
    FUNC(A, FA);
    A⇒X; 若 FA=0 则转 OUT 否;
    FUNC(B, FB);
    B⇒X; 若 FB=0 则转 OUT 否;
  ITRT: (A+B)/2⇒X; FUNC(X, FX);
    若 §ABS(X)<1 则 §ABS(B-A)⇒DX
      否 §ABS((B-A)/X)⇒DX;
    若 DX<EPS 则转 OUT 否;
    若 FX=0 则转 OUT 否; §SIGN(FX)⇒SF;
    若 SF*FA<0 则始 X⇒B; FX⇒FB; 转 ITRT 终否;
```

```
    若 SF*FB<0 则始 X⇒A; FX⇒FA; 转 ITRT 终否;
  OUT:
终;
```

二、线性分式插值法(求一个实零点)程序

HYPE($x1$, $x2$, $ep1$, $ep2$, $\max k$, x, f, Dx, *FUNC*, *FAIL*)

使用说明

过程 *HYPE* 是应用线性分式插值法(双曲插值法)计算函数 $f(x)$ 的一个实零点。

输入参数:

$x1$, $x2$——零点的两个初始近似值;

$ep1$, $ep2$——是允许误差,当相邻两次近似值 x^{k-1}, x^k 之差 $Dx<ep1$ 时,或者当函数值 $|f(x^k)|<ep2$ 时,以 x^k 作为所求之零点.此处,

$$Dx=\begin{cases}|x^k-x^{k-1}| & 当|x^k|<1\\ |x^k-x^{k-1}|/|x^k| & 当|x^k|\geqslant 1\end{cases}$$

$\max k$——允许的最大迭代次数;

FAIL——非正常出口,当迭代次数 $>\max k$ 时,将转向 *FAIL*[1];

FUNC——计算函数值 $f(x)$ 的过程。须有过程说明: 过程 *FUNC*(x, f); 值 x; 简变 f; 始…终;

输出参数:

x——零点计算值;

f——相应的函数值, $f=f(x)$;

Dx——误差。

程序

```
过程 HYPE(X1, X2, EP1, EP2, MAXK, X, F, DX, FUNC, FAIL);
  值 X1, X2, EP1, EP2, MAXK;
  简变 X, F, DX;
  过程 FUNC;
  开关 FAIL;
始  简变 F1, F2, LMD, DLT, EP3, EP5, SUP, M, MU, MUF1, MUF2, MUF3, D, K;
    2↑(-MAXP+2)⇒EP3;
   注 1 {MAXP 是机器数的最大阶码数,使用时需填入。在109(2)机上, MAXP=31}
    2↑((MAXP-1)/2-1)⇒SUP; 1/SUP⇒EP5; 0⇒K; 1⇒DX; FUNC(X1, F1);
      FUNC(X2, F2); -0.5⇒LMD;
  HYP1: (X1+X2)/2⇒X; FUNC(X, F);
    若§ABS(F)<EP2 则转 OUT 否;
  ITRT: 若§ABS(F2-F1)<EP3 则转 ITR2 否;
    若§ABS(F-F1)<EP3 则转 ITR2 否;
    若§ABS(F-F2)<EP3 则转 ITR2 否;
```

```
    若§ABS(F1)<§ABS(F2)则 F2⇒M 否 F1⇒M;
    若§ABS(M)<§ABS(F)则 F⇒M 否;
    1⇒MU; §ABS(M)⇒M;
    若 2*M<EP3 则转 ITR1 否;
    若 M≤SUP 则若 0.5≤M 则转 ITR1 否始
  DMU1: 2*M⇒M; 若 1≤M 则转 ITR1
                  否 2*MU⇒MU; 转 DMU1 终
    否始
  DMU2: MU/2⇒MU; M/2⇒M;
    若 M≤SUP 则否转 DMU2;
      终;
  ITR1: MU*F1⇒MUF1; MU*F2⇒MUF2;
    MU*F⇒MUF3; 1+LMD⇒DLT;
    MUF3*(DLT*MUF2-LMD*MUF1)-MUF3*MUF1⇒M;
    MUF3*(DLT*MUF1-LMD*MUF2)-MUF2*MUF1⇒D;
    若§ABS(D)<EP3 则转 ITR2 否;
    若§ABS(D)<§ABS(M)
    则若§ABS(D/M)<EP5 则转 ITR2 否
    否;
    M/D⇒LMD; 转 ITR3;
  ITR2: 1⇒LMD;
  ITR3: LMD*(X-X2)⇒DX;
    X2⇒X1; F2⇒F1; X⇒X2; F⇒F2; X+DX⇒X;
    若§ABS(X)<1 则否 DX/X⇒DX;
    K+1⇒K;
    FUNC(X, F);
    若§ABS(F)<EP2 则转 OUT 否;
    若§ABS(DX)<EP1 则转 OUT 否
    若 K<MAXK 则转 ITRT 否转 FAIL[1];
  OUT:
终;
```

三、线性分式插值法(求区间上全部单零点)程序
HPBL(a, b, h, ep1, ep2, x, f, FUNC, TRAT)

使用说明

过程 *HPBL* 是应用线性分式插值法(双曲插值法)和区间分割法，确定函数 $f(x)$ 在区间 $[a, b]$ 上的全部单零点。

输入参数：

a, b——定解区间的端点;

h——相邻两零点之间距的下限估值，对于间距小于 h 的零点，过程就有可能忽略;

$ep1$, $ep2$——允许误差,详见过程 $HYPE$ 之说明;

$FUNC$——计算函数值 $f(x)$ 的过程。需有过程说明: **过程** $FUNC(x, f)$; **值** x; **简变** f; **始…终**;

$TRAT$——是供用户处理和加工零点 x 的,本过程每求出一零点后便供过程 $TRAT$ 处理。

输出参数:

x——零点计算值;

f——相应的函数值, $f=f(x)$;

印刷控制:

当 #1 尾部第 1 位为 1 时,印出每次迭代值:

k(迭代次数), x^k, $f(x^k)$, Dx^k;

当 #1 尾部第 2 位为 1 时,印出每个零点值:

r(零点序号), k(迭代次数), x, $f(x)$, Dx。

程序

```
过程 HPBL(A, B, H, EP1, EP2, X, F, FUNC, TRAT);
  值 A, B, H, EP1, EP2;
  简变 X, F;
  过程 FUNC, TRAT;
始  简变 R, K, DX, I, LX, RX, LF, RF, S;
    若 H=0 则转 OUT 否;
    0⇒I; 1⇒R; A⇒RX;
  CUT: 若 B≤RX 则转 OUT 否;
    RX⇒LX; RF⇒LF; A+I*H⇒RX;
    若 B<RX 则 B⇒RX 否;
    FUNC(RX, RF); I+1⇒I;
    若 RF=0 则始
            RX⇒X; RF⇒F; 0⇒S; RX+H/10⇒RX; FUNC(RX, RF);
            转 FDRT
               终否;
    若 I=1 则否
    若 §SIGN(LF)*RF<0
    则始 1⇒S;
        若 #1∧ 字 2=字 2
注 2  {控制印刷的条件可按机器和使用要求更改之}
        则始  印数 0, LX, LF, H;
              印数 0, RX, RF, H;
```

```
        终否;
      转 FDRT
    终
  否; 转 CUT;
FDRT: 始
  场 XX[1:R]; 开关 SWIT[FAIL];
过程 HYPE(X1, X2, F1, F2, X, F, SWIT);
  简变 X1, X2, F1, F2, X, F;
  开关 SWIT;
  始  简变 F3, LMD, DLT, EP3, EP5, SUP, M, MU, FX, MUF1, MUF2, MUF3,
    D;
    过程 FUNP(T, F, FR);
      值 T; 简变 F, FR;
    始  简变 D;
      FUNC(T, F); 1⇒D;
      对于 I=1 到 R-1 步长 1 执行
        D*(T-XX[I])⇒D;
      F/D⇒FR;
      若 #1∧ 字2=字2
      则印数 K, T, FR, DX 否; 1⇒FX;
      若 §ABS(F)<EP2
      则若 §ABS(FR)<EP2 则 0⇒FX 否
      否;
    终;
  2↑(-MAXP+2)⇒EP3;
  注 {见过程 HYPE 之注 1。}
  2↑((MAXP-1)/2-1)⇒SUP; 1/SUP⇒EP5; 0⇒K; -0.5⇒LMD; (X1+X2)/2⇒
    X; FUNP(X, F, F3);
  若 FX=0 则转 OUT1 否;
ITRT: 若 §ABS(F2-F1)<EP3 则转 ITR2 否;
  若 §ABS(F3-F1)<EP3 则转 ITR2 否;
  若 §ABS(F3-F2)<EP3 则转 ITR2 否;
  若 §ABS(F1)<§ABS(F2) 则 F2⇒M 否 F1⇒M;
  若 §ABS(M)<§ABS(F3) 则 F3⇒M 否;
  1⇒MU; §ABS(M)⇒M;
  若 2*M<EP3 则转 ITR1 否;
  若 M≤SUP 则若 0.5≤M 则转 ITR1 否始
DMU1: 2*M⇒M;
    若 1≤M 则转 ITR1 否 2*MU⇒MU;
```

```
    转 DMU1
  否始
DMU2: MU/2⇒MU; M/2⇒M;
    若 M≤SUP 则否转 DMU2;
    终;
ITR1: MU*F1⇒MUF1; MU*F2⇒MUF2;
  MU*F3⇒MUF3;
  1+LMD⇒DLT;
  MUF3*(DLT*MUF2-LMD*MUF1)-MUF3*MUF1⇒M;
  MUF3*(DLT*MUF1-LMD*MUF2)-MUF2*MUF1⇒D;
  若§ABS(D)<EP3 则转 ITR2 否;
  若§ABS(D)<§ABS(M)
  则若§ABS(D/M)<EP5 则转 ITR2 否
  否;
  M/D⇒LMD;
  转 ITR3;
ITR2: 1⇒LMD;
ITR3: LMD*(X-X2)⇒DX; X2⇒X1; F2⇒F1; X⇒X2; F3⇒F2; X+DX⇒X;
  若§ABS(X)<1 则否 DX/X⇒DX; K+1⇒K;
  FUNP(X, F, F3);
  若 FX=0 则转 OUT1 否;
  若§ABS(DX)<EP1
  则转 OUT1
  否若 K<40 则转 ITRT 否转 SWIT[1];
OUT1: 终;
过程 HITL(A, B, FA, FB, X, F);
  简变 A, B, FA, FB, X, F;
  始  简变 SF;
ITRT: K+1⇒K; (A+B)/2⇒X; FUNC(X, F); B-A⇒DX;
  若§ABS(X)<1 则否 DX/X⇒DX;
  若 #1∧ 字 2=字 2
  则印数 K, X, F, DX 否;
  若§ABS(DX)<EP1 则转 OUT2 否;
  若§ABS(F)<EP2 则转 OUT2 否;
  §SIGN(F)⇒SF;
  若 SF*FA<0 则始 X⇒B; F⇒FB; 转 ITRT 终否;
  若 SF*FB<0 则始 X⇒A; F⇒FA; 转 ITRT 终否;
OUT2: 终;
  若 S=0 则转 PRIN 否
```

```
    HYPE(LX, RX, LF, RF, X, F, SWIT);
    转 PRIN;
  FAIL: HITL(LX, RX, LF, RF, X, F);
  PRIN: X⇒XX[R];
    若 #1∧ 字 4=字 4
    则印数 R, K, X, F, DX 否
    TRAT;
    R+1⇒R; 转 CUT;
  终;
  OUT:
终;
```

四、求函数零点(实或复的)的牛顿法程序

$NWTN(RZ0, IZ0, ep1, ep2, \max k, RZ, IZ, f, DZ, FUNC, FAIL)$

使用说明

过程 $NWTN$ 是应用牛顿迭代法求函数 $f(z)$ 的一个零点(实或复的)。

输入参数:

$RZ0$, $IZ0$——分别是零点初始近似的实、虚部;

$ep1$, $ep2$——允许误差,当相邻两次近似值 z^{k-1} 和 z^k 之差 $Dz<ep1$, 或者当 $f(z^k)$ 的模值 $\|f(z^k)\|<ep2$ 时,便以 z^k 作为所求之根值,此处

$$DZ=\begin{cases}\|z^k-z^{k-1}\| & \text{当 } \|z^k\|<1\\ \|z^k-z^{k-1}\|/\|z^k\| & \text{当 } \|z^k\|\geqslant 1\end{cases}$$

$\max k$——允许的最大迭代次数;

$FAIL$——非正常出口,当迭代次数 $>\max k$ 时,将转向 $FAIL[1]$;

$FUNC$——是计算函数值 $f(z)=Rf+jIf$ 和导数值 $f'(z)=Rf1+jIf1$ 的过程。过程需有说明: 过程 $FUNC(RZ, IZ, Rf, If, Rf1, If1)$; 值 RZ, IZ; 简变 $Rf, If, Rf1, If1$; 始…终;

输出参数:

RZ, IZ——分别是零点的实、虚部;

f——相应的函数值之模 $f=\|f(z)\|$;

DZ——误差。

印刷控制:

当 #1 尾部第 1 位为 1 时,将印出每次迭代值:

k(迭代次数), RZ^k, IZ^k, $Rf(Z^k)$, $If(Z^k)$。

程序

```
过程 NWTN(RZ0, IZ0, EP1, EP2, MAXK, RZ, IZ, F, DZ, FUNC, FAIL);
  值 RZ0, IZ0, EP1, EP2, MAXK;
  简变 RZ, IZ, F, DZ;
```

```
过程 FUNC;
开关 FAIL;
始  简变 RF, IF, RF1, IF1, K, RH, IH, M, M1, EP3, SUP, MU;
   0⇒K; 2↑(-MAXP+2)⇒EP3;
  注 {MAXP 是机器数的最大阶码数,使用时须填入。在109(乙)机上, MAXP=31。}
   2↑((MAXP-1)/2-1)⇒SUP; RZ0⇒RZ; IZ0⇒IZ; 0.1⇒RH;
 ITRT: FUNC(RZ, IZ, RF, IF, RF1, IF1);
   若 #1∧ 字2=字2
  注 {印刷控制的条件,可按机器及需要更改之。}
   则印数 K, RZ, IZ, RF, IF, DZ 否;
   §ABS(RF)+§ABS(IF)⇒F;
   §ABS(RF1)+§ABS(IF1)⇒M1;
   若 F<EP2 则转 OUT 否
   若 F≤M1 则否若 M1/F<EP3 则转 ITR1 否;
   若 §ABS(IF)+§ABS(IF1)=0
   则始 RF/RF1⇒RH; 0⇒IH; 转 ITR1 终否;
   0⇒M;
   对 T=RF, IF, RF1, IF1 执行
   若 §ABS(M)<§ABS(T)
   则 §ABS(T)⇒M 否;
   1⇒MU;
   若 2*M<EP3 则转 ITR1 否;
   若 M≤SUP 则若 0.5≤M 则转 DMU3 否
                始
 DMU1: 2*M⇒M;
   若 1≤M 则转 DMU3 否 2*MU⇒MU;
   转 DMU1  终
   否始
 DMU2: MU/2⇒MU; M/2⇒M;
   若 M≤SUP 则否转 DMU2
     终;
 DMU3: MU*RF⇒RF; MU*IF⇒IF;
   MU*RF1⇒RF1; MU*IF1⇒IF1;
   RF1*RF1+IF1*IF1⇒M1;
   (RF*RF1+IF*IF1)/M1⇒RH;
   (IF*RF1-RF*IF1)/M1⇒IH;
 ITR1: RZ-RH⇒RZ; IZ-IH⇒IZ; K+1⇒K;
   §ABS(RZ)+§ABS(IZ)⇒M;
   §ABS(RH)+§ABS(IH)⇒DZ;
```

```
    若 M<1 则否 DZ/M⇒DZ;
    若 DZ<EP1 则转 OUT 否;
    若 K<MAXK 则转 ITRT 否转 FAIL[1];
  OUT:
终;
```

五、二次插值法程序

MULR(*RZ*0, *IZ*0 *n*, *ep*1, *ep*2, max *k*, *RZZ*, *IZZ*, *FUNC*, *FAIL*)

使用说明

过程 *MULR* 是应用二次插值法寻求函数 $f(z)$ 在复平面上的 n 个零点。

输入参数:

n——需求零点的个数;

*RZ*0, *IZ*0——模为最小的零点的初始近似值, *RZ*0 为实部, *IZ*0 为虚部;

*ep*1, *ep*2——允许误差,详见过程 *NWTN* 之说明;

max k——允许的最大迭代次数;

FAIL——非正常出口,当迭代次数 >max k 时,将转向 *FAIL*[1];

FUNC——是计算函数值 $f(Z)=Rf+jIf$ 的过程。过程需有说明: 过程 *FUNC*(*RZ*, *IZ*, *Rf*, *If*); 值*RZ*, *IZ*; 简变 *Rf*, *If*; 始…终;

输出参数:

RZZ, *IZZ*: n 个根值,其实部赋值于 *RZZ*, 其虚部赋值于 *IZZ*。

印刷控制:

当 #1 尾部的第 1 位为 1 时,将印出每次迭代值:

k(迭代次数), RZ^k, IZ^k, $Rf(Z^k)$, $If(Z^k)$, *DZ*;

当 #1 尾部第 2 位为 1 时,印出每个根值:

r(根序号), k(迭代次数), *RZ*, *IZ*, *Rf*(*Z*), *If*(*Z*), *DZ*。

程序

```
过程 MULR(RZ0, IZ0, N, EP1, EP2, MAXK, RZZ, IZZ, FUNC, FAIL);
  值 RZ0, IZ0, N, EP1, EP2, MAXK;
  场 RZZ, IZZ;
  过程 FUNC;
  开关 FAIL;
始  简变 R, RZ, IZ, SUP, EP4, EP5, EP6, RZ1, IZ1, RZ2, IZ2, RF, IF, RF1, IF1,
  RF2, IF2, RF3, IF3, K, RLAM, ILAM, RDEL, IDEL, RA, IA, RB, IB, RC, IC,
  MU, RH, IH, RD, ID, T, DZ;
    场 XX[0:6];
  过程 DEMU(N);
    值 N;
  始  简变 M;
```

```
  0⇒M;
  对于 I=1 到 N 步长 1 执行
    若 §ABS(XX[I])≤M 则否 §ABS(XX[I])⇒M;
  M⇒XX[0];
  1⇒MU;
  若 M<EP6 则转 DOUT 否;
  若 M≤SUP 则若 0.5≤M 则转 DOUT 否始
DMU1: 2*M⇒M;
  若 1≤M 则转 DMU3 否 2*MU⇒MU; 转 DMU1 终
  否始
DMU2: MU/2⇒MU; M/2⇒M;
  若 M≤SUP 则否转 DMU2;
    终;
DMU3: 对于 I=1 到 N 步长 1 执行
        MU*XX[I]⇒XX[I];
DOUT:
终;
过程 MTPL(RA, IA, RB, IB, RC, IC);
  值 RA, IA, RB, IB;
  简变 RC, IC;
始 RA*RB-IA*IB⇒RC;
    RA*IB+RB*IA⇒IC;
终;
过程 DVID(RA, IA, RB, IB, RC, IC);
  值 RA, IA, RB, IB;
  简变 RC, IC;
始
  RA⇒XX[1]; IA⇒XX[2]; RB⇒XX[3]; IB⇒XX[4]; DEMU(4); XX[3]*XX[3]
    +XX[4]*XX[4]⇒IC; (XX[1]*XX[3]+XX[2]*XX[4])/IC⇒RC; (XX[2]*
    XX[3]-XX[1]*XX[4])/IC⇒IC;
终;
过程 SQRT(RA, IA, RC, IC);
  值 RA, IA;
  简变 RC, IC;
始  简变 T;
  若 §ABS(RA)+§ABS(IA)<EP6
  则始 0⇒RC; 0⇒IC 终
  否始 RA⇒XX[1]; IA⇒XX[2]; DEMU(2);
  §SQRT((§ABS(RA)+§SQRT(XX[1]*XX[1]+XX[2]*XX[2]/MU)/2)⇒T;
```

```
  若 RA<0 则始 IA/(2*T)⇒RC; T⇒IC 终
  否始 T⇒RC; IA/(2*T)⇒IC 终;
    终
终;
过程 FUNR(RT, IT, RFR, IFR);
  值 RT, IT;
  简变 RFR, IFR;
始  简变 RD, ID, RE, IE;
  FUNC(RT, IT, RF, IF);
  若 R<2 则始 RF⇒RFR; IF⇒IFR; 转 FUN1 终否;
  1⇒RD; 0⇒ID;
  对于 I=1 到 R-1 步长 1 执行
    始
      MTPL(RD, ID, RT-RZZ[I], IT-IZZ[I], RE, IE); RE⇒RD; IE⇒ID;
    终;
  DVID(RF, IF, RD, ID, RFR, IFR);
FUN1: 若 #1∧ 字 2=字 2
 注 {见过程 HPBL 的注 2}
  则印数 K, RT, IT, RFR, IFR 否;
  若 §ABS(RF)+§ABS(IF)<EP2
  则若 §ABS(RFR)+§ABS(IFR)<EP2
    则始 RT⇒RZ; IT⇒IZ; 转 SOUT 终
    否
  否;
终;
  2↑((MAXP-1)/2-1)⇒SUP;
 注 {见过程 HYPE 的注 1}
  2↑(-MAXP+2)⇒EP4; EP4/2⇒EP6; 2↑(-10)⇒EP5; 0⇒RZZ; 0⇒IZZ;
    1⇒R;
FDRT: 0⇒K;
  RZ0⇒RZ; IZ0⇒IZ;
  若 §ABS(RZ0)+§ABS(IZ0)<EP6
  则始 -0.25⇒RZ1; 0⇒IZ1; 0.25⇒RZ2; 0⇒IZ2 终
  否始 0.75*RZ0⇒RZ1; 0.75*IZ0⇒IZ1; 1.25*RZ0⇒RZ2; 1.25*IZ0⇒IZ2
    终;
  FUNR(RZ1, IZ1, RF1, IF1);
  FUNR(RZ2, IZ2, RF2, IF2);
  FUNR(RZ, IZ, RF3, IF3);
  -0.5⇒RLAM; 0⇒ILAM;
```

```
ITRT: 1+RLAM⇒RDEL; ILAM⇒IDEL;
  MTPL(RF1, IF1, RLAM, ILAM, RB, IB);
  MTPL(RF2, IF2, RDEL, IDEL, RC, IC);
  MTPL(RB-RC+RF3, IB-IC+IF3, RLAM, ILAM, RA, IA);
  MTPL(RF1, IF1, RLAM*RLAM-ILAM*ILAM, 2*RLAM*ILAM, RB, IB);
  MTPL(RF2, IF2, RDEL*RDEL-IDEL*IDEL, 2*RDEL*IDEL, RC, IC);
  MTPL(RF3, IF3, RLAM+RDEL, ILAM+IDEL, RH, IH);
  RB-RC+RH⇒RB; IB-IC+IH⇒IB;
  MTPL(RF3, IF3, RDEL, IDEL, RC, IC);
  RA⇒XX[1]; RB⇒XX[2]; RC⇒XX[3];
  IA⇒XX[4]; IB⇒XX[5]; IC⇒XX[6];
  DEMU(6);
  若 XX[0]<EP4 则转 TRAT 否;
  XX[1]⇒RA; XX[2]⇒RB; XX[3]⇒RC;
  XX[4]⇒IA; XX[5]⇒IB; XX[6]⇒IC;
  MTPL(RA, IA, RC, IC, RH, IH);
  SQRT(RB*RB-IB*IB-4*RH, 2*RB*IB-4*IH, RH, IH);
  若 §ABS(RB+RH)+§ABS(IB+IH)<§ABS(RB-RH)+§ABS(IB-IH)
  则始 RB-RH⇒RD; IB-IH⇒ID 终
  否始 RB+RH⇒RD; IB+IH⇒ID 终;
  若 §ABS(RD)+§ABS(ID)<EP4
  则转 TRAT 否;
  2*RC⇒RC; 2*IC⇒IC;
  §ABS(RC)+§ABS(IC)⇒RH;
  §ABS(RD)+§ABS(ID)⇒T;
  若 T<RH 则若 T/RH<EP5 则转 TRAT 否否;
  DVID(-RC, -IC, RD, ID, RLAM, ILAM);
  转 ITR1;
TRAT: 1.1⇒RLAM; 0⇒ILAM;
ITR1: RZ2⇒RZ1; IZ2⇒IZ1;
  RF2⇒RF1; IF2⇒IF1;
  RZ⇒RZ2; IZ⇒IZ2;
  RF3⇒RF2; IF3⇒IF2;
  K+1⇒K;
  MTPL(RLAM, ILAM, RZ2-RZ1, IZ2-IZ1, RH, IH);
ITR2: RZ2+RH⇒RZ; IZ2+IH⇒IZ;
  §ABS(RH)+§ABS(IH)⇒DZ;
  §ABS(RZ)+§ABS(IZ)⇒T;
  若 T<C 则否 DZ/T⇒DZ;
```

```
    FUNR(RZ, IZ, RF3, IF3);
    若 §ABS(RF3)+§ABS(IF3)≤10*(§ABS, (RF2)+§ABS(IF2))则否
    始 RLAM/2⇒RLAM; ILAM/2⇒ILAM;
      RH/2⇒RH; IH/2⇒IH; 转 ITR2
    终;
    若 DZ<EP1 则转 SOUT 否;
    若 K<MAXK 则转 ITRT 否转 FAIL[1]
  SOUT: 若 #1∧ 字 4=字 4
    则印数 R, K, RZ, IZ, RF, IF, DZ 否;
    RZ⇒RZZ[R]; IZ⇒IZZ[R];
    1.005*RZ⇒RZ0; 1.005*IZ⇒IZ0;
    R+1⇒R;
    若 R≤N 则转 FDRT 否;
终;
```

六、解非线性方程组的牛顿法程序

$SNWT(\boldsymbol{X}0, n, ep1, ep2, \max k, \boldsymbol{X}, \boldsymbol{F}, \boldsymbol{JF}, Dx, FUNC, FAIL)$

使用说明

过程 $SNWT$ 是用牛顿迭代法求非线性方程组 $\boldsymbol{F}(\boldsymbol{X})=0$ 的一组解，此处 $\boldsymbol{X}$ 和 $\boldsymbol{F}(\boldsymbol{X})$ 都是 n 元向量：$\boldsymbol{F}=\{f_1, f_2, \cdots, f_n\}$，$\boldsymbol{X}=\{x_1, \cdots, x_n\}$。

输入参数：

n——方程个数；

$\boldsymbol{X}0$——解的初始近似值(n 元向量)；

$ep1$, $ep2$——允许误差，当相邻两次近似值 $\boldsymbol{X}^{k-1}, \boldsymbol{X}^k$ 之差 $DX<ep1$，或者当 $\|\boldsymbol{F}(\boldsymbol{X}^k)\|<ep2$ 时，以 $\boldsymbol{X}^k$ 作为所求的解 $\boldsymbol{X}$，此处 $DX=\max\limits_i\{\Delta x_i^k\}$

$$\Delta x_i^k=\begin{cases}|x_i^k-x_i^{k-1}| & \text{当 } |x_i^k|<1\\ |x_i^k-x_i^{k-1}|/|x_i^k| & \text{当 } |x_i^k|\geqslant 1\end{cases}$$

$$\|\boldsymbol{F}(\boldsymbol{x}^k)\|=\max_i\{|f_i(\boldsymbol{X}^k)|\};$$

$\max k$——允许的最大迭代次数；

$FAIL$——非正常出口，当迭代次数 $>\max k$ 时，转向 $FAIL[1]$；

$FUNC$——是计算相应于 $\boldsymbol{X}$ 的函数值 $\boldsymbol{F}=\{f_1, \cdots, f_n\}$ 和偏导数值矩阵

$$\boldsymbol{JF}=\begin{pmatrix}\dfrac{\partial f_1}{\partial x_1}, & \cdots, & \dfrac{\partial f_1}{\partial x_n}\\ \cdots\cdots & & \\ \dfrac{\partial f_n}{\partial x_1}, & \cdots, & \dfrac{\partial f_n}{\partial x_n}\end{pmatrix}$$

的过程，过程需有说明。

$\boldsymbol{JF}$——由 $\boldsymbol{X}$ 处的 $\boldsymbol{F}(\boldsymbol{X})$ 之偏导数值所组成的 $n\times n$ 阶矩阵；

输出参数:

$\boldsymbol{X}$——解;
$\boldsymbol{F}$——$\boldsymbol{X}$ 相应的函数值; } n 元向量

DX——误差。

印刷控制:

当 #1 尾部第 1 位为 1 时, 印出每次迭代值:

k(迭代次数), $\boldsymbol{X}$, $\boldsymbol{F}$, DX。

程序

```
过程 SNWT(X0, N, EP1, EP2, MAXK, X, F, JF, DX, FUNC, FAIL);
  值 N, EP1, EP2, MAXK;
  场 X0, X, F, JF;
  简变 DX;
  过程 FUNC;
  开关 FAIL;
始  简变 K, MF, D;
  过程 ELMT;
  始  简变 AM, II, JJ, AA;
    对于 R=1 到 N 步长 1 执行
      始 0⇒AM;
        对于 J=1 到 N 步长 1 执行
          对于 I=R 到 N 步长 1 执行
            若 §ABS(AM)<§ABS(JF[I, J])
            则始 I⇒II; J⇒JJ; JF[I, J]⇒AM 终
            否;
      若 §ABS(AM)<2↑(-MAXP+1)
   注 {见过程 HYPE 的注 1}
      则转 FAIL[1]否;
      对于 J=1 到 N 步长 1 执行
        始 JF[II, J]⇒AA; JF[R, J]⇒JF[II, J]; AA/AM⇒JF[R, J]
        终;
      F[II]⇒AA; F[R]⇒F[II]; AA/AM⇒F[R];
      对于 I=1 到 N 步长 1 执行
        若 I=R 则否始
          JF[I, JJ]⇒AA;
          对于 J=1 到 N 步长 1 执行
            JF[I, J]-AA*JF[R, J]⇒JF[I, J]; F[I]-AA*F[R]⇒F[I];
                    终;
      终;
    对于 J=1 到 N 步长 1 执行
```

```
     始 1⇒II;
   ELM1: 若 §ABS(JF[II, J])<0.9
     则始 II+1⇒II; 转 ELM1 终
     否始对于 K=1 到 N 步长 1 执行
           始 JF[II, K]⇒AA; JF[J, K]⇒JF[II, K];
             AA⇒JF[J, K]
           终;
         F[II]⇒AA; F[J]⇒F[II]; AA⇒F[J];
         终;
     终
  终;
   0⇒K; X0⇒X; EP1⇒DX;
  ITRT: FUNC;
    若 #1∧ 字 2=字 2
   注 {见过程 HPBL 的注 2}
    则印数 K, X, F, DX 否;
    0⇒MF;
    对于 I=1 到 N 步长 1 执行
      若 §ABS(F[I])<MF
      则否 §ABS(F[I])⇒MF;
    若 MF<EP2 则转 OUT 否;
    若 DX<EP1
    则转 OUT
    否若 K<MAXK 则否转 FAIL[1];
    ELMT;
    0⇒DX;
    对于 I=1 到 N 步长 1 执行
      始 F[I]⇒D; X[I]-D⇒X[I];
      若 §ABS(X[I])<1 则否 D/X[I]⇒D;
      若 §ABS(D)≤DX 则否 §ABS(D)⇒DX
      终;
    K+1⇒K;
    转 ITRT;
  OUT:
终;
```

七、解非线性方程组的最速下降法和牛顿迭代法程序

$DSNT(\boldsymbol{X}0, n, ep1, ep2, \max k, \boldsymbol{X}, \boldsymbol{F}, \boldsymbol{JF}, Dx, FUNC, FAIL)$

使用说明

过程 $DSNT$ 是应用最速下降法和牛顿迭代法求非线性方程组 $\boldsymbol{F}(\boldsymbol{X})=0$ 的一组解。此处，$\boldsymbol{F}(\boldsymbol{X})=\{f_1, f_2, \cdots, f_n\}$，$\boldsymbol{X}=\{x_1, x_2, \cdots, x_n\}$是 n 元向量。

输入参数:

n——方程个数;

$ep1$, $ep2$——允许误差,详见过程 $SNWT$ 之说明;

$\boldsymbol{X}0$——解之初始近似值(n 元向量);

$\max k$——允许的最大迭代次数;

$FAIL$——非正常出口，当迭代次数 $>\max k$ 或遇其他计算失败情况时将转向 $FAIL[1]$;

$\boldsymbol{JF}$——由 $\boldsymbol{X}$ 处的 $\boldsymbol{F}(\boldsymbol{X})$ 之偏导数值所组成的 $n\times n$ 阶矩阵:

$$\boldsymbol{JF}=\begin{pmatrix} \frac{\partial f_1}{\partial x_1}, & \cdots, & \frac{\partial f_1}{\partial x_n} \\ \cdots\cdots & & \\ \frac{\partial f_n}{\partial x_1}, & \cdots, & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

$FUNC$——计算 $\boldsymbol{X}$ 处的函数值 $\boldsymbol{F}(\boldsymbol{X})$ 和偏导数值 $\boldsymbol{JF}$ 的过程。过程需有说明: **过程** $FUNC(S)$; **值** S; **始**…**终**;。当 $S=1$ 时,仅要求计算 $\boldsymbol{F}(\boldsymbol{X})$，赋值于 $\boldsymbol{F}$; 当 $S=2$ 时,仅要求计算 $\boldsymbol{JF}$; 当 S 为其他值时,要求同时计算 $\boldsymbol{F}$ 和 $\boldsymbol{JF}$;

输出参数:

$\boldsymbol{X}$——解;
$\boldsymbol{F}$——$\boldsymbol{X}$ 处的函数值; } n 元向量

Dx——误差。

印刷控制:

当 #1 的第 1 位为 1 时,印出每次迭代值; 当 #1 的第 2 位为 1 时,印出结果,印出形式均为:

k(迭代次数), $\boldsymbol{X}$, $\boldsymbol{F}$, Dx。

程序

```
过程 DSNT(X0, N, EP1, EP2, MAXK, X, F, JF, DX, FUNC, FAIL);
  值 N, EP1, EP2, MAXK;
  场 X0, X, F, JF;
  简变 DX;
  过程 FUNC;
  开关 FAIL;
始
```

```
简变 K;
场 G[1:N];
过程 SPDT;
始
  简变 MG, FE, FE1, EP3, LMD, JG, P, ALFA;
  10⇒DX;
  2↑(-MAXP+1)⇒EP3;
 注 {见过程 HYPE 之注 1}
  FUNC(1); 0⇒FE;
  对于 I=1 到 N 步长 1 执行
    FE+§ABS(F[I])⇒FE;
SITR: FUNC(2);
  若 #1∧ 字 2=字 2
 注 {见过程 HPBL 之注 2}
  则印数 K, X, F, DX 否;
  若 FE<0.1 则转 SOUT 否;
  若 DX<1 则转 SOUT 否;
  若 K<MAXK 则否转 SOUT;
  0⇒G; 0⇒MG;
  对于 I=1 到 N 步长 1 执行
  始对于 J=1 到 N 步长 1 执行
      G[I]+JF[J, I]*F[J]⇒G[I];
    若 MG<§ABS(G[I])
    则 §ABS(G[I])⇒MG
    否;
  终;
  若 MG<EP3 则转 SOUT 否;
  对于 I=1 到 N 步长 1 执行
    G[I]/MG⇒F[I]; 0⇒LMD; 0⇒JG;
  对于 I=1 到 N 步长 1 执行
  始 0⇒P;
    对于 J=1 到 N 步长 1 执行
      P+JF[I, J]*F[J]⇒P; JG+P*P⇒JG; LMD+F[I]*F[I]⇒LMD;
  终;
  LMD/JG⇒LMD; 0⇒DX;
  对于 I=1 到 N 步长 1 执行
  始 LMD*G[I]⇒G[I];
    若 §ABS(X[I])<1
    则 G[I]⇒P
```

```
    否 G[I]/X[I]⇒P;
    若 §ABS(P)≤DX 则否 §ABS(P)⇒DX;
  终;
  -1⇒ALFA;
SIT1: 对于 I=1 到 N 步长 1 执行
      X[I]+ALFA*G[I]⇒X[I];
  FUNC(1);
  0⇒FE1;
  对于 I=1 到 N 步长 1 执行
    FE1+§ABS(F[I])⇒FE1;
  若 FE1≤FE
  则转 SIT2
  否 §ABS(ALFA)/2⇒ALFA;
  若 ALFA<0.1 则否转 SIT1;
SIT2: FE1⇒FE; K+1⇒K;
  转 SITR;
SOUT:
终;
过程 SNWT;
始
  简变 MF, D;
  过程 ELMT;
 注 {过程内容见前面过程 SNWT 中的同名过程,使用时请插入}
  EP1⇒DX;
ITRT: FUNC(0);
  若 #1∧ 字 2=字 2
 注 {见过程 HPBL 之注 2}
  则印数 K, X, F, DX
  否;
  0⇒MF;
  对于 I=1 到 N 步长 1 执行
    若 §ABS(F[I])<MF
    则否 §ABS(F[I])⇒MF;
  若 MF<EP2 则转 NOUT 否;
  若 DX<EP1 则转 NOUT 否;
  若 K<MAXK 则否转 FAIL[1];
  ELMT;
  0⇒DX;
  对于 I=1 到 N 步长 1 执行
```

```
    始 F[I]⇒D; X[I]-D⇒X[I];
      若 §ABS(X[I])<1 则否 D/X[I]⇒D;
      若 §ABS(D)≤DX 则否 §ABS(D)⇒DX;
    终;
    K+1⇒K;
    转 ITRT;
  NOUT;
  终;
    0⇒K; X0⇒X;
    SPDT;
    SNWT;
    若 #1∧ 字4=字4
  注 {见过程 HPBL 之注 2}
    则印数 K, X, F, DX
    否;
终;
```

八、DFP 方法程序

$VMTC(X0, n, ep1, ep2, \max k, X, F, JF, Dx, FUNC, FAIL)$

使用说明

过程 $VMTC$ 是用 DFP 方法求非线性方程组 $\boldsymbol{F}(\boldsymbol{X})=0$ 的一组解，此处 $\boldsymbol{F}(\boldsymbol{X})$，$\boldsymbol{X}$ 均是向量：$\boldsymbol{F}(\boldsymbol{X})=\{f_1, f_2, \cdots, f_n\}$，$\boldsymbol{X}=\{x_1, x_2, \cdots, x_n\}$。

输入参数：

n——方程个数；

$ep1$, $ep2$——允许误差，详见过程 $SNWT$ 之说明；

$X0$——解之初始近似值(n 元向量)；

$\max k$——允许的最大迭代次数；

$FAIL$——非正常出口，当迭代次数 $>\max k$ 时，将转向 $FAIL[1]$；

JF——$\boldsymbol{X}$ 处的 $\boldsymbol{F}(\boldsymbol{X})$ 之偏导数值所组成的 $n\times n$ 阶矩阵：

$$\boldsymbol{JF}=\begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}, & \cdots, & \dfrac{\partial f_1}{\partial x_n} \\ \cdots\cdots & & \\ \dfrac{\partial f_n}{\partial x_1}, & \cdots, & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix}$$

$FUNC$——计算 $\boldsymbol{X}$ 处的函数值 F 和偏导数值 JF 的过程，需有过程说明；

输出参数：

X——解；
F——$\boldsymbol{X}$ 处的函数值 $F=\boldsymbol{F}(\boldsymbol{X})$；
（以上为 n 元向量）

Dx——误差。

印刷控制:

当 #1 的第 1 位为 1 时,印出每次迭代值;当 #1 的第 2 位为 1 时,印出结果。印出形式均为:

$$k(\text{迭代次数}), X, F, Dx$$

程序

```
过程 VMTC(X0, N, EP1, EP2, MAXK, X, F, JF, DX, FUNC, FAIL);
  值 N, EP1, EP2, MAXK;
  场 X0, X, F, JF;
  简变 DX;
  过程 FUNC;
  开关 FAIL;
始
    简变 ALFA, K, L, MF, MF0, EP3, D;
    场 G[1:N, 1:N];
      X1, P, GRAD, P1, G1[1:N];
  过程 CG(X1, GRAD);
    场 X1, GRAD;
  始
    X⇒GRAD; X1⇒X;
    FUNC;
    GRAD⇒X; 0⇒GRAD; 0⇒MF;
    对于 I=1 到 N 步长 1 执行
    始对于 J=1 到 N 步长 1 执行
        2*JF[J, I]*F[J]+GRAD[I]⇒GRAD[I];
        若 §ABS(F[I])≤MF
        则否 §ABS(F[I])⇒MF;
    终
  终;
  过程 CP;
  始
    0⇒P;
    对于 I=1 到 N 步长 1 执行
      对于 J=1 到 N 步长 1 执行
        P[I]-G[I, J]*GRAD[J]⇒P[I];
  终;
  过程 MUTG;
  始
    简变 D1, D2, MP, MDG;
    0⇒MP; 0⇒MDG;
```

```
    对于 I=1 到 N 步长 1 执行
    始若 MP<§ABS(P1[I])
      则 §ABS(P1[I])⇒MP 否;
      若 MDG<§ABS(G1[I])
      则 §ABS(G1[I])⇒MDG 否;
    终;
    若 MP≤EP3 则转 MOUT 否;
    若 MDG≤EP3 则转 MOUT 否;
    对于 I=1 到 N 步长 1 执行
    始 P1[I]/MP⇒P1[I];
      G1[I]/MDG⇒G1[I];
    终;
    0⇒D1; 0⇒D2; 0⇒X1;
    对于 I=1 到 N 步长 1 执行
    始对于 J=1 到 N 步长 1 执行
        X1[I]+G[I, J]*G1[J]⇒X1[I];
      D2+G1[I]*X1[I]⇒D2; D1+P1[I]*G1[I]⇒D1;
    终;
    对于 I=1 到 N 步长 1 执行
      对于 J=1 到 N 步长 1 执行
        G[I, J]+(P1[I]*P1[J]/D1)*(MP/MDG)-X1[I]*X1[J]/D2⇒G[I, J];
MOUT:
终;
过程 CALF;
始
    简变 LMD, T, KK, H, Z, W, Y, Y1, Y0, Y01;
    MF⇒MF0; MF⇒W; 0⇒Y0; 0⇒Y01; 0⇒H;
    对于 I=1 到 N 步长 1 执行
    始 F[I]/MF0⇒T; Y0+T*T⇒Y0; P[I]/MF0⇒T; T⇒P1[I]; H+T*T⇒H;
      Y01+(GRAD[I]/MF0)*T⇒Y01;
    终;
    1/(MF0*§SQRT(H))⇒H; -2*Y0/Y01⇒KK;
    若 §ABS(KK)≤EP3
    则否若 KK<H 则 KK⇒H 否;
    H⇒LMD;
CAL1: 对于 I=1 到 N 步长 1 执行
          X[I]+LMD*P[I]⇒X1[I];
    CG(X1, G1);
    若 MF≤EP2
```

```
  则始 LMD⇒ALFA; 转 COUT 终否;
  0⇒Y; 0⇒Y1;
  对于 I=1 到 N 步长 1 执行
  始 F[I]/MF⇒T;
    Y+T*T⇒Y; Y1+(G1[I]/MF)*P1[I]⇒Y1;
  终;
  若 §SIGN(Y01)*Y1≤0
  则否始 Y⇒Y0; Y1⇒Y01;
        MF⇒MF0; 2*LMD⇒LMD;
        转 CAL1
        终;
  (W+MF0+MF)/3⇒Z;
  W/Z⇒W; MF0/Z⇒T; MF/Z⇒Z;
  T*Y0*T⇒Y0; T*Y01*W⇒Y01;
  Z*Y*Z⇒Y; Z*Y1*W⇒Y1;
  6*(Y0-Y)/LMD+Y01+Y1⇒Z;
  §SQRT(Z*Z-Y01*Y1)⇒W;
  (1-(Y1+W-Z)/(2*(Y1-Y01+2*W)))*LMD⇒ALFA;
COUT:
终;
  2↑(-MAXP+1)⇒EP3;
 注 {见过程 HYPE 之注 1}
  EP1⇒DX; 0⇒K; X0⇒X; 0⇒L;
LITR: 0⇒G;
  对于 I=1 到 N 步长 1 执行
    1⇒G[I, I]; CG(X, GRAD); CP;
ITRT: 若 #1∧ 字 2=字 2
 注 {见过程 HPBL 之注 2}
      则印数 K, X, F, DX 否;
  若 §ABS(MF)<EP2
  则转 OUT 否;
  若 DX<EP1 则转 OUT 否;
  CALF; 0⇒DX;
  对于 I=1 到 N 步长 1 执行
  始 ALFA*P[I]⇒P1[I]; X[I]+P1[I]⇒X[I]; §ABS(P1[I])⇒D;
    若 §ABS(X[I])<1
    则否 D/§ABS(X[I])⇒D;
    若 D≤DX 则否 D⇒DX;
  终;
```

```
K+1⇒K; L+1⇒L;
若 DX<0.001
则否若 N≤L
    则始 0⇒L; 转 LITR 终
    否;
GRAD⇒G1; CG(X, GRAD);
对于 I=1 到 N 步长 1 执行
GRAD[I]-G1[I]⇒G1[I]; MUTG; CP;
若 K<MAXK
则转 ITRT
否转 FAIL[1];
OUT: 若 #1∧ 字4=字4
        注 {见过程 HPBL 之注 2}
    则印数 K, X, F, DX
    否;
终;
```

参考资料

[1] Ostrowski, A. M., "Solution of Equations and Systems of Equations", Academic Press, New York, 1960.

[2] Muller, D. E., "A method for solving algebraic equations using an automatic computer", Math. Tables Aids Comput., Vol. 10(1956), pp. 208~215.

[3] Jarratt, P. and Nudds, D., "The use of rational functions in the iterative solution of equations on a digital computer", The Computer Journal, Vol. 8, No. 1, (1965).

[4] Wilkinson, J. H., "The evaluation of the zeros of ill-conditioned polynomials." Numer. Math., Vol. 1 (1959), No. 3.

[5] Davidon, W. C., "Variable metric method for minimization", A. E. C. Research and Development Report ANL-5990 (1959).

[6] Fletcher, R. and Powell, M. J. D., "A rapidly convergent desent method for minimization", Computer Journal, Vol. 6, No. 2 (1963), pp. 163~168.

[7] Bard, Y., "On a numerical" instability of Davidon-like methods" Mathematics of Computation, Vol. 22, No. 103, (1968), pp. 665~666.

[8] Powell, M. T. D., "Recent advances in Unconstrained optimization", A. E. R. E. Report TPE 430, (1970).

第十章　代数特征值问题的解法

§10.1　引　　言

在工程实践中，经常遇到振动问题以及相应的代数特征值问题。大型桥梁或建筑物的振动问题、机械和机件的振动问题、飞机机翼的颤振问题、无线电工学及光学系统的电磁振荡问题、调节系统和随动系统中的自振问题以及声学和超声系统的振动问题等，这些都是明显的例子。尽管各种振动过程的物理本质是互不相同的，但处理它们的数学方法却是一致的，这一点有利于使用计算机解决问题。同时，工程实践中所提出的振动问题往往比较复杂，人工去进行解算通常是不可能的，而计算机却可以有效地解决它们。实际上，我国许多大型工程的设计中，都使用了计算机来解决振动问题。

通常，使用计算机求解振动问题要经过如下几个步骤：

(1) 针对要求解的物理振动问题，选择描述它的数学模型(有时也把一个数学模型称为一个振动系统)，并确定该数学模型中的有关常数。

(2) 从数学模型出发，推导出便于在计算机上求解的数学方程式。

(3) 针对该数学方程式的特点和所使用的计算机的情况，选择恰当的数值解法编制程序上机计算。

前面两步可以称为问题的提法，后一步即具体的解法。本章只讨论与 2、3 两步有关的问题。讨论与 2 有关问题的目的，也仅在于说明实践中所提出的代数特征值问题的某些特点。

工程实践中的振动问题，通常是有阻尼出现的。但有时阻尼的影响很小，为了简便起见可将其忽略，这样得出的振动系统称为无阻尼系统，否则叫做阻尼系统。此外，振动系统还可以分为连续系统与离散系统两类。所谓离散系统，是指该系统具有有限的自由度数，其状态可用有限个广义坐标 $q_1, q_2, \cdots, q_n$ 来描述。这类系统所遵守的微分方程通常是一个常微分方程组。连续系统则与之不同，它的自由度数不是有限的，描述它们的微分方程通常是偏微分方程组及其初始或边界条件。虽然工程实践中的振动问题严格说来都是连续系统，但在许多情况下，用有限个自由度已经可以相当准确地进行描述。特别是在计算机上求解时，连续系统也必须化为近似的离散系统，所以这里主要讨论离散系统的处理办法。对于连续系统的离散化问题，本书第十三、十四章有详细论述，这里仅概略说明离散化的基本步骤，以便阐明振动问题的某些特点。还应指出，本章讨论的振动问题仅限于围绕平衡位置的小振动问题，其相应的微分方程组是线性常系数方程组(对连续系统，其系数可以是自变数的函数)，这样的系统称为线性系统，在计算机上求解时，一般归结为代数特征值问题。对于非线性系统的求解，往往需要直接去解描述它们的非线性微分方程组。关于这类问题在计算机上的解法，可以参阅本书第十一章至第十三章的有关部分。

首先叙述振动问题的提法(§10.2)，然后叙述相应的代数特征值问题在计算机上的解法(§10.3)。为使用方便，对于 §10.3 所讨论的主要方法附有用算法语言编写的计算机程序，这些程序均在机器上进行过计算，读者可以直接使用或者转换为具体机器上的相应算法语

言程序来使用。

§10.2 振动问题的提法

本节讨论如何推导振动系统的微分方程和怎样将其求解化为代数特征值问题。同时，以此为例来说明实践中所提出的代数特征值问题的主要特点。

10.2.1 有限自由度系统

由于各种物理振动问题的处理方法是大同小异的，因此这里只讨论动力学系统的小振动问题。假定该系统有 n 个自由度，其位置可用 n 个广义坐标 $q_i(i=1, 2, \cdots, n)$ 描述，同时，$q_1=q_2=\cdots=q_n=0$ 为系统的平衡位置。那么，系统围绕这一平衡位置的小振动问题归结为确定广义坐标 q_i 与时间 t 的依赖关系：

$$q_i=q_i(t) \quad (i=1, 2, \cdots, n)$$

如果作用于该系统的力可分为两部分，一部分力的相应势能函数 U 仅为广义坐标 q_i 的函数：$U=U(q_1, q_2, \cdots, q_n)$；另一部分为与时间 t 有关的广义力 $Q_i(t)$，$(i=1, 2, \cdots, n)$。势能函数 U 可在平衡位置 $q_i=0$ 附近展开：

$$U(q_1, q_2, \cdots, q_n)=U_0+\sum_{i=1}^{n}\alpha_i q_i+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}q_iq_j+\cdots$$

由于 $q_i=0$ 系平衡位置，势能应取稳定值，所以 $\alpha_i=0$。此外，平衡位置的势能 U_0 究竟等于多少是无关紧要的，可以将其取为零。再用小振动的假定，可将上式中三次以上的各项略去。于是势能可以表成：

$$U=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}q_iq_j$$

或者

$$U=\frac{1}{2}q^T\cdot A\cdot q \tag{10.2.1}$$

其中

$$a_{ji}=a_{ij}=\left(\frac{\partial U}{\partial q_i\cdot\partial q_j}\right)_{q_i=0}$$

为常系数；$\boldsymbol{q}^T=\{q_1, q_2, \cdots, q_n\}$为行向量；矩阵 $\boldsymbol{A}=[a_{ij}]$ 为对称矩阵，如果平衡位置 $q_i=0$ 是稳定的，则 $\boldsymbol{A}$ 应为正定矩阵。

系统的动能一般由下式表达：

$$T=\frac{1}{2}\sum_{i=1}^{n}m_iv_i^2$$

质量 m_i 为当量质量，相应的速度 v_i 为 $\dot{q}_i=\frac{dq_i}{dt}$ 的函数：$v_i=f_i(\dot{q}_1, \dot{q}_2, \cdots, \dot{q}_n)$。同样对 f_i 进行展开，并忽略 $\dot{q}_i$ 二次以上各项即得：

$$v_i=\sum_{j=1}^{n}l_{ij}\dot{q}_j \qquad l_{ij}=\frac{\partial f_i}{\partial \dot{q}_j}$$

这样，动能 T 可以表为：

$$T=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ij}\dot{q}_i\dot{q}_j$$

或者

$$T=\frac{1}{2}\dot{q}^T\cdot B\cdot\dot{q} \tag{10.2.2}$$

其中

$$b_{ji}=b_{ij}=\sum_{k=1}^{n}m_k l_{ki} l_{kj}$$

为常系数；矩阵 $\boldsymbol{B}=[b_{ij}]$ 为对称矩阵。同时，由于动能总是恒正的(仅当 $\dot{q}=0$ 时，T 为零)，所以，$\boldsymbol{B}$ 为正定矩阵。

有了动能 T 与势能 U 的表达式后，运动方程式就可从最小作用量原理(Hamilton 原理)推导出来。最小作用量原理可表述为：

$$\delta_W=\int_{t_0}^{t_1}\left[\delta(T-U)+\sum_{i=1}^{n}Q_i\delta q_i\right]dt=0$$

其中，$\delta q_i(t)$ 应满足条件：

$$\delta q_i(t)\,|_{t=t_0}=\delta q_i(t)\,|_{t=t_1}=0 \quad (i=1,\ 2,\ \cdots,\ n)$$

将 T 及 U 的表达式代入后即得

$$\delta_W=\int_{t_0}^{t_1}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}[b_{ij}\dot{q}_j(\delta\dot{q}_i)-a_{ij}q_j(\delta q_i)]+\sum_{i=1}^{n}Q_i\cdot\delta q_i\right\}dt$$

注意到 $\delta\dot{q}_i=(\delta q_i)^{\cdot}$，并对方括号中第一项进行分部积分即得

$$\delta_W=\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ij}\dot{q}_j(\delta q_i)\Big|_{t_0}^{t_1}-\int_{t_0}^{t_1}\left\{\sum_{i=1}^{n}\left[\sum_{j=1}^{n}(b_{ij}\ddot{q}_j+a_{ij}q_j)-Q_i\right]\delta q_i\right\}dt$$

从 δq_i 在 $t=t_0$ 及 $t=t_1$ 为零的条件得知上式第一项应为零。此外，由于 δq_i 是独立的，故得到

$$\sum_{j=1}^{n}(b_{ij}\ddot{q}_j+a_{ij}q_j)-Q_i(t)=0 \quad (i=1,\ 2,\ \cdots,\ n)$$

或者写成矩阵形式

$$\boldsymbol{B}\ddot{\boldsymbol{q}}+\boldsymbol{A}\boldsymbol{q}=\boldsymbol{Q}(t)$$

其中，$\boldsymbol{Q}(t)=(Q_1(t),\ Q_2(t),\ \cdots,\ Q_n(t))^T$，这就是要求的无阻尼时系统的运动方程式。

上述方程式就是所谓的动力学系统的拉格朗日(Lagrange)方程式：

$$\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_i}\right)-\frac{\partial T}{\partial q_i}+\frac{\partial U}{\partial q_i}=Q_i(t) \quad (i=1,\ 2,\ \cdots,\ n)$$

处理动力学系统时，可直接将动能及势能表达式代入上式来得出运动方程式，这与应用最小作用量原理是等效的。但后者的变分处理办法在其它问题中也经常使用。

如果系统在有阻尼的介质中运动，则将有与速度成比例的阻尼 $\boldsymbol{R}=\{R_i\}$ 作用于系统的各质点上，这样的力 $\{R_i\}$ 通常可用所谓耗散函数 D 来表示。在小振动的情况下，它可表为

$$D=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}\dot{q}_i\dot{q}_j \quad 或者 \quad D=\frac{1}{2}\dot{q}^T\cdot C\cdot\dot{q}$$

$$R_i=\frac{-\partial D}{\partial\dot{q}_i}=-\sum_{j=1}^{n}C_{ij}\dot{q}_j \quad (i=1,\ 2,\ \cdots,\ n) \tag{10.2.3}$$

其中，$C_{ij}=C_{ji}$ 为阻尼系数；矩阵 $\boldsymbol{c}=[C_{ij}]$ 为对称矩阵；$\boldsymbol{R}=\{R_i\}$ 为相应的阻尼力。

有阻尼时的拉格朗日方程式为：

$$\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_i}\right)-\frac{\partial T}{\partial q_i}+\frac{\partial U}{\partial q_i}+\frac{\partial D}{\partial\dot{q}_i}=Q_i(t) \quad (i=1,\ 2,\ \cdots,\ n) \tag{10.2.4}$$

把 T、U、D 的表达式代入其中，就得到有阻尼时系统小振动的运动方程式

$$\sum_{j=1}^{n}(b_{ij}\ddot{q}_j+C_{ij}\dot{q}_j+a_{ij}q_j)=Q_i(t)\quad(i=1,\ 2,\ \cdots,\ n)$$

或者

$$\boldsymbol{B}\ddot{\boldsymbol{q}}+\boldsymbol{c}\dot{\boldsymbol{q}}+\boldsymbol{A}\boldsymbol{q}=\boldsymbol{Q}(t)\tag{10.2.5}$$

其中 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 均为对称矩阵。

10.2.2 连续系统

对于连续系统，可以用一个近似于它的有限自由度系统来代替，把问题转化为有限自由度系统的求解。这个方法叫做"离散化"。这样求得的解是原来问题的近似解。只要"离散化"的方法是合理的，求得的近似解一般也是可用的。关于这方面的详细内容可参看本书第十四章。这里仅叙述一下其简单的处理原则，以便阐明连续系统振动问题的某些特点。

在空间直角坐标系$\{x,\ y,\ z\}$中，连续体的位移分量可用三个位移函数表示：

$$\begin{cases}u=U(x,\ y,\ z,\ t)\\ v=V(x,\ y,\ z,\ t)\\ w=W(x,\ y,\ z,\ t)\end{cases}$$

按照通常有限元法的思想，将连续体分割为若干子区域 D_k，在每个子区域 D_k 上分别对 u、v、w 定义相应的插值函数 u_k、v_k、w_k（跨过相邻子区域 D_k，D_j 的边界时，插值函数 u_k，u_j；v_k，v_j；w_k，w_j 等应满足一定的连续性条件。此外，每个插值函数还应满足所谓"常应变"条件等等）。将这些插值函数拼接起来，就得到整个区域 D 上的插值函数。这样，整个区域上位移函数的插值函数可以表为：

$$u=\sum_k U_k\quad v=\sum_k V_k\quad w=\sum_k W_k$$

其中

$$U_k=\begin{cases}u_k & (x,\ y,\ z)\in D_k\\ 0 & \text{其它}\end{cases}$$

这些插值函数一般将由若干个参数或称广义坐标 q_1，q_2，$\cdots$，q_n（例如，节点位移及其导数等）所确定，广义坐标 q_i 则仅为时间 t 的函数。于是，用上述插值函数代替位移函数后，连续体的位移分量就可以近似地表为

$$u=u(x,\ y,\ z,\ q_1,\ q_2,\ \cdots,\ q_n)$$

$$v=v(x,\ y,\ z,\ q_1,\ q_2,\ \cdots,\ q_n)$$

$$w=w(x,\ y,\ z,\ q_1,\ q_2,\ \cdots,\ q_n)$$

这里 u、v、w 均为 x、y、z 及 q_i 的已知函数。如果求出了广义坐标 $q_i(t)$，那么，连续体内任意点的位移即可近似地得到。这样，确定连续体位移的问题，也就近似地化成了如何求得广义坐标q_i的有限自由度问题。只要合理地对区域进行分割及适当地选取插值函数，一般来说，当子区域的体积无限缩小时，有限自由度问题的解将收敛于连续体问题的解。所以，这种近似在一定条件下总是合理的。

下面，来导出广义坐标 $q_i(t)$ 所应满足的运动方程式。仍然假定只考虑围绕平衡位置 $q_1=q_2=\cdots=q_n=0$ 的小振动问题。由于这个条件，位移函数还可进一步简化为

$$\begin{cases} u=\phi_0(x,\ y,\ z)+\sum\limits_{i=1}^{n}\phi_i(x,\ y,\ z)q_i \\ v=\psi_0(x,\ y,\ z)+\sum\limits_{i=1}^{n}\psi_i(x,\ y,\ z)q_i \\ w=\chi_0(x,\ y,\ z)+\sum\limits_{i=1}^{n}\chi_i(x,\ y,\ z)q_i \end{cases} \tag{10.2.6}$$

其中

$$\phi_i(x,\ y,\ z)=\left.\frac{\partial u}{\partial q_i}\right|_{q_1=q_2=\cdots=q_n=0}$$

等等。即是说，由于 q_i 本身应为小量，故可在展开式中忽略其二次以上各项。

连续体动能 T 的表达式为：

$$T=\frac{1}{2}\iiint\limits_V \rho(x,\ y,\ z)(\dot{u}^2+\dot{v}^2+\dot{w}^2)dV$$

将其中的位移函数代之以插值函数来求出相应有限自由度系统的动能。由于

$$\begin{cases} \dot{u}=\sum\limits_i \dfrac{\partial u}{\partial q_i}\dot{q}_i=\sum\limits_i \phi_i(x,\ y,\ z)\dot{q}_i \\ \dot{v}=\sum\limits_i \dfrac{\partial v}{\partial q_i}\dot{q}_i=\sum\limits_i \psi_i(x,\ y,\ z)\dot{q}_i \\ \dot{w}=\sum\limits_i \dfrac{\partial w}{\partial q_i}\dot{q}_i=\sum\limits_i \chi_i(x,\ y,\ z)\dot{q}_i \end{cases} \tag{10.2.7}$$

故有

$$\begin{aligned} T&=\frac{1}{2}\iiint \rho(x,\ y,\ z)\cdot[(\sum_i \phi_i\dot{q}_i)^2+(\sum_i \psi_i\dot{q}_i)^2+(\sum_i \chi_i\dot{q}_i)^2]dV \\ &=\frac{1}{2}\sum_{i,j}\left[\iiint\limits_V \rho(x,\ y,\ z)\cdot(\phi_i\phi_j+\psi_i\psi_j+\chi_i\chi_j)dV\right]\dot{q}_i\dot{q}_j \\ &=\frac{1}{2}\sum_{i,j} b_{ij}\dot{q}_i\dot{q}_j=\frac{1}{2}\dot{\boldsymbol{q}}^T\boldsymbol{B}\dot{\boldsymbol{q}} \end{aligned}$$

其中

$$b_{ij}=b_{ji}=\iiint\limits_V \rho(x,\ y,\ z)(\phi_i\phi_j+\psi_i\psi_j+\chi_i\chi_j)dV \tag{10.2.8}$$

矩阵 $\boldsymbol{B}=[b_{ij}]$ 为对称正定矩阵，通常称之为质量矩阵。

连续体势能 U 的表达式为：

$$U=\iiint\limits_V W(u,\ v,\ w,\ u_x,\ v_x,\ w_x,\ \cdots,\ w_z)dV$$

其中 W 为应变能密度函数，其表达式随问题而异(参考资料[20]中，列有常见的各种 W 的表达式)。一般来说，采用虎克定律，W 将为位移 $u,\ v,\ w$ 及其对空间变量一阶导数 $u_x,\ v_x,\cdots,\ w_z$ 的二次泛函数。将插值函数(10.2.6)式代入其中，容易得知 W 将为广义坐标 q_i 的二次函数。同样，U 亦为 q_i 的二次函数：

$$U=a_0+\sum_i a_iq_i+\frac{1}{2}\sum_{i,j}a_{ij}q_iq_j$$

式中，系数 a_i 及 a_{ij} 为已知函数 $\phi_i,\ \psi_i,\ \chi_i$ 及其导数的某些二次式的积分值。如果平衡位置 $q_i=0$ 的势能取为零，同时由于该处势能取稳定值，可以推知：

$$a_0=a_1=a_2=\cdots=a_n=0$$

于是，势能 U 可表为：

$$U=\frac{1}{2}\sum_{i,j}a_{i,j}q_iq_j$$
$$=\frac{1}{2}\boldsymbol{q}^T\boldsymbol{A}\boldsymbol{q}$$

其中 $a_{ij}=a_{ji}$ 为常数，矩阵$[a_{ij}]=\boldsymbol{A}$ 为对称矩阵，通常称之为刚度矩阵。一般情况下，$\boldsymbol{A}$ 是正定或正半定矩阵。

如果存在与速度成比例的阻尼，则对于任意微体元 dV，单位时间内耗散能量的一半为：

$$\frac{1}{2}(C_1(x,\ y,\ z)\dot{u}^2+C_2(x,\ y,\ z)\dot{v}^2+C_3(x,\ y,\ z)\dot{w}^2)dV$$

其中 $C_i(x,\ y,\ z)$为相应的阻尼系数。这样，耗散函数 D 的表达式为：

$$D=\frac{1}{2}\iiint_V[C_1(x,\ y,\ z)\dot{u}^2+C_2(x,\ y,\ z)\dot{v}^2+C_3(x,\ y,\ z)\dot{w}^2]dV$$

将(10.2.7)代入上式即得：

$$D=\frac{1}{2}\sum_{i,j}\left\{\iiint_V[C_1(x,\ y,\ z)\phi_i\phi_j+C_2(x,\ y,z)\psi_i\psi_j+C_3(x,\ y,\ z)\chi_i\chi_j]dV\right\}\dot{q}_i\dot{q}_j$$

令

$$C_{ij}=C_{ji}=\iiint_V[C_1(x,\ y,\ z)\phi_i\phi_j+C_2(x,\ y,\ z)\psi_i\psi_j+C_3(x,\ y,\ z)\chi_i\chi_j]dV \quad (10.2.9)$$

便有：

$$D=\frac{1}{2}\sum_{i,j}C_{ij}\dot{q}_i\dot{q}_j=\frac{1}{2}\boldsymbol{q}^T\boldsymbol{C}\dot{\boldsymbol{q}}$$

矩阵 $\boldsymbol{C}=[C_{ij}]$为对称矩阵，通常称之为阻尼矩阵。

作用于连续体的外力，一般有体力分量 X、Y、Z 及面力分量 X_s、Y_s、Z_s。因而在任意微小位移 δ_u, δ_v, δ_w 上，外力所作的功 W_e 可表为：

$$W_e=\iiint_V(X\cdot\delta u+Y\cdot\delta v+Z\cdot\delta w)dV+\iint_S(X_s\delta u+Y_s\delta v+Z_s\delta w)dS$$

将(10.2.6)代入其中即有：

$$W_e=\iiint_V\left[\sum_{i=1}^{n}(X\phi_i+Y\psi_i+Z\chi_i)\delta q_i\right]dV+\iint_S\left[\sum_{i=1}^{n}(X_s\phi_i+Y_s\psi_i+Z_s\chi_i)\delta q_i\right]dS$$
$$=\sum_{i=1}^{n}\left[\iiint_V(X\phi_i+Y\psi_i+Z\chi_i)dV+\iint_S(X_s\phi_i+Y_s\psi_i+Z_s\chi_i)dS\right]\delta q_i$$

所以，与广义坐标 q_i 相应的广义力 Q_i 可以写为：

$$Q_i=\iiint_V(X\phi_i+Y\psi_i+Z\chi_i)dV+\iint_S(X_s\phi_i+Y_s\psi_i+Z_s\chi_i)dS$$

由于 X, Y, $\cdots$, Z_s 为 x, y, z 和时间 t 的函数，所以 Q_i 仅为时间 t 的函数。

有了动能、势能，耗散函数及广义力的表达式后，完全仿照有限自由度系统的处理办法，将它们代入拉格朗日方程式(10.2.4)，同样，可以得出相应的运动方程为：

$$\boldsymbol{B}\ddot{\boldsymbol{q}}+\boldsymbol{C}\dot{\boldsymbol{q}}+\boldsymbol{A}\boldsymbol{q}=\boldsymbol{Q}(t)$$

其中 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 亦均为对称矩阵。

10.2.3 化为代数特征值问题

从前面的讨论知道，有限自由度系统或连续系统振动问题的求解，均归结为解一组线性常系数的常微分方程组(10.2.5)。

实践中有时需要求出某一时间间隔$[t_0, t_1]$内系统的位移变化情况，这类问题即所谓瞬态模拟问题，只需从t_0时刻的初始条件出发，对方程(10.2.5)进行逐步地数值积分，即可求得解答。

另一类问题中需要求出的量是系统的本征振动频率与振型等等，这类问题将归结为代数特征值问题。例如，对于无阻尼自由振动问题，方程式(10.2.5)变为：

$$\boldsymbol{B}\ddot{\boldsymbol{q}}+\boldsymbol{A}\boldsymbol{q}=0$$

其解答可以表为如下形式：

$$\boldsymbol{q}=\boldsymbol{x}\cos\omega t \qquad 或 \qquad q_i=x_i\cos\omega t \quad (i=1,\ 2,\ \cdots,\ n)$$

其中，向量$\boldsymbol{x}$称为系统本征振动的振型向量；ω称为本征振动的振频。将其代入方程式，并约去其公因子$\cos\omega t$即得：

$$\sum_{j=1}^{n}(a_{ij}-\omega^2 b_{ij})x_j=0 \quad (i=1,\ 2,\ \cdots,\ n)$$

令$\lambda=\omega^2$，并写成矩阵形式便有：

$$\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{B}\boldsymbol{x} \tag{10.2.10}$$

这即是通常的广义代数特征值问题。其中矩阵$\boldsymbol{A}$与$\boldsymbol{B}$就是系统势能及动能表达式中的系数矩阵，它们都是对称矩阵，$\boldsymbol{B}$同时还是正定的。作为这个广义代数特征值问题的解，共有n个实的非负广义特征值λ_i及相应的广义特征向量$\boldsymbol{x}^{(i)}$。所以，无阻尼自由振动问题一般有n个本征振动频率$\omega_i=\sqrt{\lambda_i}$及相应的振型向量$\boldsymbol{x}^{(i)}$。

适当选取广义坐标，有时可使矩阵$\boldsymbol{B}$变为单位矩阵，这样，问题就直接化为普通代数特征值问题：

$$\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x} \tag{10.2.11}$$

其中$\boldsymbol{A}$为对称矩阵。

有阻尼自由振动问题中，方程式(10.2.5)变为：

$$\boldsymbol{B}\ddot{\boldsymbol{q}}+\boldsymbol{C}\dot{\boldsymbol{q}}+\boldsymbol{A}\boldsymbol{q}=0$$

这是一个线性齐次常系数的常微分方程组，其解答具有如下形式：

$$\boldsymbol{q}=\boldsymbol{x}e^{\omega t}$$

代入方程并约去公因子$e^{\omega t}$即得：

$$(\omega^2\boldsymbol{B}+\omega\boldsymbol{C}+\boldsymbol{A})\boldsymbol{x}=0 \tag{10.2.12}$$

这仍然是一个广义代数特征值问题，作为其解共有$2n$个ω_i值及其相应的向量$\boldsymbol{x}^{(i)}$。此时，ω_i一般为复数，其实部为振幅衰减因子，虚部为振频，$\boldsymbol{x}^{(i)}$仍为振型向量。由于(10.2.12)与如下普通代数特征值问题等价(即其解答ω及$\boldsymbol{x}$满足下式)：

$$\begin{pmatrix} O, & I \\ -\boldsymbol{B}^{-1}\boldsymbol{A}, & -\boldsymbol{B}^{-1}\boldsymbol{C} \end{pmatrix}\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}=\omega\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \tag{10.2.13}$$

所以，求解(10.2.12)可以化为非对称矩阵的普通代数特征值问题。

综上所述，振动问题的求解最终将归结为求解代数特征值问题(10.2.10)、(10.2.11)、

(10.2.12)、(10.2.13)。其中 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 均为对称矩阵，$\boldsymbol{B}$ 为正定矩阵，有时 $\boldsymbol{A}$ 亦为正定的。

除振动问题外，工程实践中还有许多问题的求解也归结为上述类型的代数特征值问题。例如，弹性结构的稳定性分析问题，各种数学物理方程的特征值问题以及某些相关分析问题、图论问题等等。在计算机上求解代数特征值问题往往是解决这些问题的关键步骤之一。从下节起，将开始讨论一些常用的求解代数特征值问题的方法。在转入下节的讨论之前，再对特征值问题的所谓“稳定性”概念作一些介绍。

如果代数特征值问题的精确解答，非常灵敏于问题原始数据的微小变化，则称此特征值问题为“不稳定的”或“病态的”。问题中的矩阵就叫作对于计算特征值来说的“病态”矩阵。

由于工程实际问题的解答一般不是非常灵敏于其各个物理参数的微小变化的，所以，将它们提成代数特征值问题时，也应该是稳定的。特别是物理参数的测量以及特征值问题中矩阵元素的形成等等，都有不可避免的误差。如果特征值问题不稳定，当然不能期望通过求解它来得出实际问题的合理解答。因此，要求形成特征值问题时，应得出稳定的特征值问题是很重要的。

从代数特征值问题的摄动分析中得知，如果问题中的矩阵是对称(或更一般的，是正规的)矩阵，那么，这一问题必定是稳定的。即是说如果对称矩阵 $\boldsymbol{A}$ 的元素有微小变化 Δa_{ij}，其特征值的变化 $\Delta\lambda$ 和特征向量的变化 $\Delta \boldsymbol{x}_i$ 将分别满足如下不等式(参阅[9]第二章)：

$$|\Delta\lambda| \leqslant \|\Delta \boldsymbol{A}\|$$

$$\|\Delta \boldsymbol{x}_i\| \leqslant \left(\sum_{j\neq i} \frac{1}{|\lambda_j-\lambda_i|}\right)\cdot \|\Delta \boldsymbol{A}\| \quad (i=1,\ 2,\ \cdots,\ n)$$

从这些不等式得知，矩阵 $\boldsymbol{A}$ 的元素有微小变化时，其特征值也只有微小变化。如果没有重特征值，其特征向量亦只有微小变化(如果有重特征值，上式说明其相应特征向量可能有很大变化，这一点我们应予以注意)。所以，只要把实际问题提成为对称矩阵的特征值问题，并保证原始数据的一定精确度，求得的结果一般就应该是可靠的(这里暂且假定计算方法的精确度是不成问题的)。此外，由于对称矩阵特征值问题也有较多有效的计算方法可供使用，因而，在形成代数特征值问题时，应该尽可能地将其提成为对称矩阵的特征值问题。这就是在前面关于问题提法的讨论中强调矩阵对称性的理由之一。

§10.3　代数特征值问题的数值解法

10.3.1　概述

从前面的讨论知道，振动问题及其它某些工程实践问题的求解，最终归结为求某些矩阵的(广义的)特征值和特征向量，即所谓(广义)代数特征值问题。本节讨论一些在计算机上求解代数特征值问题的最常用的方法。这里先讨论普通的代数特征值问题，并给出一些计算机程序，最后再简单讨论一下广义代数特征值问题。

通常有两类方法求解代数特征值问题。一类是从原始矩阵出发，用有限个相似变换将其化为便于求出特征多项式的形式，然后求特征多项式的根作为矩阵的特征值。由于多项式特别是高阶多项式的求根问题有其特有的困难之处，并且重根的计算往往精度较低等等，

故通过特征多项式来求矩阵特征值的方法，从数值计算的观点来看，不是一个好的方法。此外，从原始矩阵求得特征多项式的系数的过程，也往往对于舍入误差的影响异常灵敏，计算过程中的舍入误差，常常使最终结果的精确度受到很大影响。所以，尽管这类方法有工作量小和应用范围广的优点，其大多数在计算机上目前已很少使用。仅是其中使用正交相似变换进行化简，然后用不直接通过特征多项式求根的一些办法，目前还经常使用。本节中要讨论的用镜像映射矩阵将对称矩阵化为三对角型的方法就属于这一类。

另一类方法是迭代法（严格说来，前一类方法也是迭代法，因为矩阵的特征值是不可能由其元素经有限次算术运算得出的。所以，求特征值的任何算法都是迭代性质的。主要是为了区分方便，通常把前一类方法称之为直接法）。它不通过特征多项式，而是将特征值及特征向量作为一个无限序列的极限来求得。这类方法对舍入误差的影响有较强的稳定性，但其工作量较大。由于计算机的工作速度较高，这一缺点能得到一定程度的弥补，所以，计算机上通常使用迭代法。

实践中所提出的特征值问题其类型是多种多样的，要求也各不相同，必须针对问题特点进行具体分析，选择适当的方法。例如，实践中最常遇到的对称矩阵的特征值问题，如果其阶数不高（例如几十阶或上百阶），就可以用旋转法或化为三对角线型的方法有效地求得其全部特征值及特征向量。如果只求其部分特征值或特征向量，则用化为三对角线型的方法更适宜。对于带状的对称矩阵，则应使用适合于带型特点的化为三对角型的方法来进一步提高能够处理的阶数和节省工作量。对于阶数很高的“稀疏”矩阵，例如上千阶的绝大多数元素为零的矩阵或带型矩阵等，则用同时迭代法（或反同时迭代法）求其按模最大（或最小）的几个特征值及相应特征向量较合适。对于非对称的任意矩阵，如果阶数不高，可用广义旋转法求其特征值和特征向量或用更为有效的 $\boldsymbol{QR}$ 方法求其特征值，然后用反幂法求其相应特征向量。对于广义特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$，如果阶数不高，用平方根法分解矩阵 $\boldsymbol{B}$，然后化为对称阵的特征值问题是行之有效的方法。如果 $\boldsymbol{A}$、$\boldsymbol{B}$ 均为高阶稀疏矩阵或带型矩阵，则应使用反同时迭代法或施斗姆序列法求解。恰当地使用上述几种方法，就能够解决实践中所提出的一般问题，因而，本节主要讨论上述的几种方法。

除特殊说明外，总以 $\boldsymbol{A}$ 代表要讨论的 $(n\times n)$ 矩阵，并假定其元素均为实数，其特征值以按模递减次序排列为：

$$\begin{aligned}|\lambda_1|=|\lambda_2|=\cdots|\lambda_{r_1}|>|\lambda_{r_1}+1|\\ =\cdots=|\lambda_{r_2}|>\cdots>|\lambda_{r_{m-1}}+1|\\ =\cdots=|\lambda_{r_m}|\end{aligned}\tag{10.3.1}$$

其相应特征向量为 $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\cdots$, $\boldsymbol{x}_{r_m}$（$r_m=n$——矩阵 $\boldsymbol{A}$ 的阶数）。这些向量被认为是按其长度为 1 或其最大模元素为 1 进行归一化的。

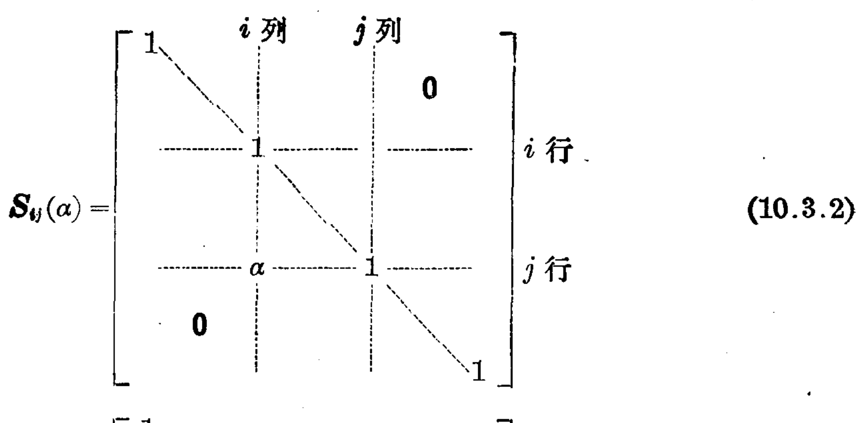
10.3.2 几种变换矩阵及其特性

先介绍几种简单的变换矩阵，这些矩阵以后要经常用到。

（一）初等变换矩阵

将矩阵 $\boldsymbol{A}$ 的第 i 行（j 列）乘以一个实数 α 后加到其第 j 行（i 列）上去或将矩阵 $\boldsymbol{A}$ 的 i, j 两行（列）互换（$i<j$），叫做对矩阵 $\boldsymbol{A}$ 进行初等变换。它们分别相当于将矩阵 $\boldsymbol{A}$ 左乘（右乘）以如下矩阵：

$S_{ij}(\alpha)=$



(10.3.2)

$$
P_{ij}=\begin{bmatrix}
1 & & & & & & & & \\
 & \ddots & & & & & & \mathbf{0} & \\
 & & 1 & & & & & & \\
 & & & 0 & \cdots & 1 & & & \\
 & & & \vdots & 1 & & & & \\
 & & & & & \ddots & & & \\
 & & & & & 1 & & & \\
 & & & 1 & \cdots & 0 & & & \\
 & & & & & & 1 & & \\
 & \mathbf{0} & & & & & & \ddots & \\
 & & & & & & & & 1
\end{bmatrix}
\begin{matrix} \\ \\ \\ i\text{ 行} \\ \\ \\ \\ j\text{ 行} \\ \\ \\ \\ \end{matrix}
\tag{10.3.3}
$$

矩阵 $\boldsymbol{S}_{ij}(\alpha)$，$\boldsymbol{P}_{ij}$ 称为初等变换矩阵。很容易验证如下等式：

$$\boldsymbol{S}_{ij}^{-1}(\alpha)=\boldsymbol{S}_{ij}(-\alpha)$$

$$\boldsymbol{P}_{ij}^{-1}=\boldsymbol{P}_{ij}$$

若对矩阵 $\boldsymbol{A}$ 进行如下初等相似变换：

$$\boldsymbol{A}^{(1)}=\boldsymbol{S}_{ij}(\alpha)\cdot\boldsymbol{A}\cdot\boldsymbol{S}_{ij}^{-1}(\alpha) \tag{10.3.4}$$

则可以看到矩阵 $\boldsymbol{A}$ 中除第 j 行和第 i 列元素外，其它元素不变。第 j 行与第 i 列元素按下列公式计算：

$$\begin{cases} a_{jl}^{(1)}=a_{jl}+\alpha\cdot a_{il} & l\neq i \\ a_{li}^{(1)}=a_{li}-\alpha\cdot a_{lj} & l\neq j \\ a_{ji}^{(1)}=(a_{ji}+\alpha\cdot a_{ii})-\alpha\cdot(a_{jj}+\alpha a_{ij}) \end{cases} \tag{10.3.5}$$

如果我们计算一下变换前后矩阵元素平方和的变化，便可得到：

$$N^2(\boldsymbol{A})-N^2(\boldsymbol{A}^{(1)})=\sum_{k,l}a_{kl}^2-\sum_{k,l}a_{kl}^{(1)2}=2\alpha C_{ji}-a\alpha^2-b\alpha^3-c\alpha^4 \tag{10.3.6}$$

其中

$$C_{ji}=\sum_{l=1}^{n}(a_{lj}a_{li}-a_{jl}a_{il})$$

$$a=\sum_{\substack{l=1\\l\neq j}}^{n}a_{lj}^2+\sum_{\substack{l=1\\l\neq i}}^{n}a_{il}^2+(a_{jj}-a_{ii})^2-2a_{ij}a_{ji}$$

$$b=2a_{ij}(a_{jj}-a_{ii})$$

$$c=a_{ij}^2$$

（二）旋转矩阵

我们知道，将空间直角坐标系 $\{x,\ y,\ z\}$ 的两个坐标轴 x 和 y 在平面 x-y 内绕 z 轴旋转

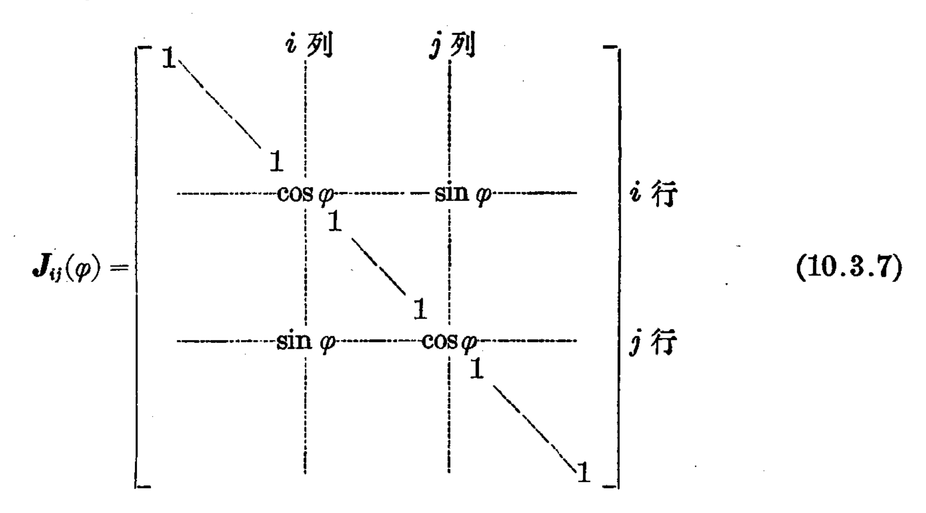
一个角度φ时，空间各点的坐标按下式进行变换：

$$\begin{Bmatrix} x \\ y \\ z \end{Bmatrix} = \begin{bmatrix} \cos\varphi, & -\sin\varphi, & 0 \\ \sin\varphi, & \cos\varphi, & 0 \\ 0, & 0, & 1 \end{bmatrix} \cdot \begin{Bmatrix} x' \\ y' \\ z' \end{Bmatrix}$$

其中变换矩阵

$$\boldsymbol{J}_{xy} = \begin{bmatrix} \cos\varphi, & -\sin\varphi, & 0 \\ \sin\varphi, & \cos\varphi, & 0 \\ 0, & 0, & 1 \end{bmatrix}$$

为三维空间内的平面旋转矩阵。如果是在n维空间中将相互正交的两个坐标轴在其所决定的平面上旋转一个角度φ，并保持正交坐标系的其它轴不动，则变换矩阵为：



(10.3.7)

矩阵$\boldsymbol{J}_{ij}(\varphi)$为$n$维空间内的平面旋转矩阵。以后，将$\boldsymbol{J}_{ij}(\varphi)$简称为平面旋转矩阵。显然，$\boldsymbol{J}_{ij}(\varphi)$是正交矩阵，即：

$$\boldsymbol{J}_{ij}(\varphi)^T \cdot \boldsymbol{J}_{ij}(\varphi) = \boldsymbol{I}$$

如果将矩阵$\boldsymbol{A}$左乘以矩阵$\boldsymbol{J}_{ij}(\varphi)^T$，则很容易看出矩阵$\boldsymbol{A}$仅有$i$, j两行元素有所变化，其余元素不变。以$\tilde{\boldsymbol{A}}=(\tilde{a}_{ij})$记为乘积矩阵，便有：

$$\begin{cases} \tilde{a}_{kl} = a_{kl} & k \neq i,\ j \\ \tilde{a}_{il} = a_{il}\cos\varphi + a_{jl}\sin\varphi & l = 1,\ 2,\ \cdots,\ n \\ \tilde{a}_{jl} = -a_{il}\sin\varphi + a_{jl}\cos\varphi & \end{cases} \tag{10.3.8}$$

从这个式子看出，适当地选择角度φ，便可以使矩阵$\tilde{\boldsymbol{A}}$第j行的任一个元素$\tilde{a}_{jl}=0$。为此，只需令：

$$\begin{cases} \cos\varphi = \dfrac{a_{il}}{\sqrt{a_{il}^2 + a_{jl}^2}} \\ \sin\varphi = \dfrac{a_{jl}}{\sqrt{a_{il}^2 + a_{jl}^2}} \end{cases} \text{若 } \sqrt{a_{il}^2 + a_{jl}^2} \neq 0 \\ \cos\varphi = 1,\ \sin\varphi = 0 \quad \text{若 } \sqrt{a_{il}^2 + a_{jl}^2} = 0 \tag{10.3.9}$$

将$\tilde{\boldsymbol{A}}$再乘以矩阵$\boldsymbol{J}_{ij}(\psi)$，以$\boldsymbol{A}^{(1)}=(a_{ij}^{(1)})$记乘积矩阵，可得：

$$
\begin{cases}
A^{(1)}=J_{ij}(\varphi)^T\cdot A\cdot J_{ij}(\psi) \\
a_{kl}^{(1)}=a_{kl} \qquad k\neq i,\ j, \quad l\neq i,\ j \\
a_{il}^{(1)}=a_{il}\cos\varphi+a_{jl}\sin\varphi \\
a_{jl}^{(1)}=-a_{il}\sin\varphi+a_{jl}\cos\varphi \qquad l\neq i,\ j \\
a_{li}^{(1)}=a_{li}\cos\psi+a_{lj}\sin\psi \\
a_{lj}^{(1)}=-a_{li}\sin\psi+a_{lj}\cos\psi \\
a_{ii}^{(1)}=(a_{ii}\cos\varphi+a_{ji}\sin\varphi)\cos\psi+(a_{ij}\cos\varphi+a_{jj}\sin\varphi)\sin\psi \\
a_{jj}^{(1)}=-(-a_{ii}\sin\varphi+a_{ji}\cos\varphi)\sin\psi+(-a_{ij}\sin\varphi+a_{jj}\cos\varphi)\cos\psi \\
a_{ij}^{(1)}=-(a_{ii}\cos\varphi+a_{ji}\sin\varphi)\sin\psi+(a_{ij}\cos\varphi+a_{jj}\sin\varphi)\cos\psi \\
a_{ji}^{(1)}=(-a_{ii}\sin\varphi+a_{ji}\cos\varphi)\cos\psi+(-a_{ij}\sin\varphi+a_{jj}\cos\varphi)\sin\psi
\end{cases}
\tag{10.3.10}
$$

计算一下矩阵 $A^{(1)}$ 非对角线元素的平方和，可以得出：

$$
\begin{aligned}
\sum_{k\neq l}a_{kl}^{(1)2}=\sum_{k\neq l}a_{kl}^2-(a_{ij}^2+a_{ji}^2)+[(-a_{ii}\sin\varphi+a_{ji}\cos\varphi)\cos\psi \\
+(-a_{ij}\sin\varphi+a_{jj}\cos\varphi)\cdot\sin\psi]^2+[-(a_{ii}\cos\varphi+a_{ji}\sin\varphi)\sin\psi \\
+(a_{ij}\cos\varphi+a_{jj}\sin\varphi)\cos\psi]^2
\end{aligned}
\tag{10.3.11}
$$

如果矩阵 A 是对称的，令 $\psi=\varphi$，则很容易看出 $A^{(1)}$ 也是对称矩阵，且如下公式成立：

$$
\begin{cases}
a_{ij}^{(1)}=a_{ji}^{(1)}=-(a_{ii}\cos\varphi+a_{ji}\sin\varphi)\sin\varphi+(a_{ij}\cos\varphi+a_{jj}\sin\varphi)\cos\varphi \\
\sum\limits_{k\neq l}a_{kl}^{(1)2}=\sum\limits_{k\neq l}a_{kl}^2-2a_{ij}^2+\dfrac{1}{2}[(a_{jj}-a_{ii})\sin 2\varphi+2a_{ij}\cos 2\varphi]^2
\end{cases}
\tag{10.3.12}
$$

（三）镜像映射矩阵

镜像映射矩阵及其基本性质已在第八章第一节讨论过，这里不再重复。

代数特征值问题的讨论中，主要将使用镜像映射矩阵来作正交相似变换。下面讨论一下与此有关的问题。假定仍然采用 8.1.6 节的符号，我们来说明使用镜像映射矩阵作正交相似变换，很容易将任意矩阵 A 化为所谓海森堡(Hessenberg)型或伪三角型(矩阵 A 之元素 a_{ij} 满足条件：$a_{ij}=0(i>j+1)$者，称为上海森堡型；满足条件：$a_{ij}=0(j>i+1)$者，称为下海森堡型)。由于这一事实很有用处，我们来较详细地讨论一下。首先，我们取原始矩阵 A 的第一列(令其第一个元素为 0)：$(0,\ a_{21},\ a_{31},\ \cdots,\ a_{n1})^T$ 为第八章(8.1.29)式中的 S，取 $(0,\ 1,\ 0,\ \cdots,\ 0)^T$ 为其中的 l 来形成镜像映射矩阵 $H_{(1)}$，并对 A 作如下正交相似变换：

$$
A_2=H_{(1)}\cdot A\cdot H_{(1)}^T \tag{10.3.13}
$$

由于 $H_{(1)}^T$ 右乘 $H_{(1)}\cdot A$ 时不影响后者第一列，所以很容易得知：

$$
A_2=H_{(1)}\cdot A\cdot H_{(1)}^T=\begin{bmatrix}
\tilde a_{11} & \tilde a_{12} & \cdots\cdots & \tilde a_{1n} \\
\alpha_1 & \tilde a_{22} & & \vdots \\
0 & \tilde a_{32} & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & \tilde a_{n2} & \cdots\ \tilde a_{nn-1} & \tilde a_{nn}
\end{bmatrix}
\tag{10.3.14}
$$

变换的第二步是取矩阵 A_2 的第二列(令其前两个元素为零)：$(0,\ 0,\ \tilde a_{32},\ \cdots,\ \tilde a_{n2})^T$ 为 S，取 $l=(0,\ 0,\ 1,\ 0,\cdots,\ 0)^T$ 来形成变换矩阵 $H_{(2)}$。注意到以 $H_{(2)}$ 左乘(或右乘)A_2 时，并不改变 A_2 的前两行(或前两列)，因而，容易验证：

$$
A_3 = H_{(2)}H_{(1)}AH_{(1)}^T H_{(2)}^T = H_{(2)}H_{(1)}AH_{(1)}H_{(2)}
$$

$$
= \begin{bmatrix} \times & \cdots & \cdots & \cdots & \times \\ \alpha_1 & \times & & & \vdots \\ 0 & \alpha_2 & \times & & \vdots \\ 0 & 0 & \times & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0, & 0, & \times & \cdots & \times \end{bmatrix} \tag{10.3.15}
$$

依此类推，作 $n-2$ 次变换后，矩阵 A 将被化为下列上海森堡型矩阵：

$$
A_{n-1} = H_{(n-2)} \cdot H_{(n-3)} \cdot \cdots H_{(2)} \cdot H_{(1)} \cdot A \cdot H_{(1)} \cdot H_{(2)} \cdot \cdots \cdot H_{(n-2)}
$$

$$
= \begin{bmatrix} \times & \cdots & \cdots & \cdots & \times \\ \alpha_1 & \times & & & \vdots \\ & \alpha_2 & \times & & \vdots \\ & & \alpha_3 & \ddots & \vdots \\ \mathbf{0} & & & \ddots & \vdots \\ & & & \alpha_{n-1} & \times \end{bmatrix} \tag{10.3.16}
$$

显然，若 A 为对称矩阵，则 A_{n-1} 亦为对称矩阵，因而，其上三角部分应仅有第一次对角线元素非零，并等于 $(\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_{n-1})$。所以，此时 A 将被化为对称的三对角线型矩阵。

10.3.3 幂法及其推广

(一)幂法

幂法是通过求矩阵特征向量来求出特征值的一种迭代法。它主要是用来求矩阵按模最大的特征值和相应特征向量的。其优点是算法简单，很容易在机器上实现，对于高阶的稀疏矩阵较合适。缺点是收敛速度较慢，其有效性依赖于矩阵特征值的分布，如果按模较大的几个特征值很接近，此法用处是不大的。

幂法的基本思想是从任取的初始向量 $x^{(0)}$ 出发，用矩阵 A 逐次地乘这个向量，构成如下序列：

$$
x^{(0)},\ x^{(1)} = Ax^{(0)},\ x^{(2)} = Ax^{(1)},\ \cdots,\ x^{(k)} = Ax^{(k-1)},\ \cdots
$$

当次数 k 逐渐增大时，序列的收敛情况将与按模最大的几个特征值有密切的关系。分析这一序列的极限，即可求得按模最大的特征值与相应的特征向量。下面就来分析这一序列的收敛情况，这里，假定矩阵 A 的特征向量是完全的(即初等因子是线性的)。

初始向量 $x^{(0)}$ 可以表为矩阵特征向量 x_i 的线性组合：

$$
x^{(0)} = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{10.3.17}
$$

这样：

$$
\begin{aligned} x^{(k)} &= Ax^{(k-1)} = A^2 x^{(k-2)} = \cdots = A^k x^{(0)} \\ &= a_1 \lambda_1^k x_1 + a_2 \lambda_2^k x_2 + \cdots + a_n \lambda_n^k x_n \end{aligned}
$$

(1) 如果按模最大的特征值 λ_1 是单实根。

此时，上式可写为：

$$x^{(k)}=\lambda_1^k\cdot\left(a_1x_1+a_2\left(\frac{\lambda_2}{\lambda_1}\right)^k x_2+a_3\left(\frac{\lambda_3}{\lambda_1}\right)^k x_3+\cdots+a_n\left(\frac{\lambda_n}{\lambda_1}\right)^k x_n\right) \quad (10.3.18)$$

若 $a_1\neq0$，由上式知道当 k 充分大时，

$$x^{(k)}\sim\lambda_1^k\cdot(a_1x_1+e_k)$$

其中 e^k 为一可以忽略的向量，即 $x^{(k)}$ 与特征向量 x_1 相差一个常数因子。只需将 $x^{(k)}$ 除以其按模最大的元素进行归一化，就得到特征向量 x_1。即使 $a_1=0$，由于计算过程的舍入误差，必将引入在 x_1 方向上的微小分量，这一分量随着迭代过程的进展而逐渐居于主导地位，其收敛情况最终也将与 $a_1\neq0$ 时类似。此外，特征值 λ_1 亦很容易求得，例如，可用如下公式计算：

$$\lambda_1\sim\|x^{(k+1)}\|\big/\|x^{(k)}\| \quad (k\text{ 充分大})$$

从上述分析可以看出，幂法的收敛速度虽与初始向量 $x^{(0)}$ 的选择有关，但主要是依赖于比例 λ_2/λ_1。这一比值愈小，收敛愈快。当此值接近于1时，收敛将很慢。

实际计算中，为了避免逐次迭代向量 $x^{(k)}$ 变得很大或很小，通常在迭代的每一步以 $x^{(k)}$ 的按模最大分量 $\max(x^{(k)})$ 来除 $x^{(k)}$ 的各个分量，从而得出归一化的向量 $y^{(k)}$，并令

$$x^{(k+1)}=A\cdot y^{(k)}$$

因而，实际计算时所用公式为：

$$\begin{cases} y^{(k)}=x^{(k)}/\max(x^{(k)}) \\ x^{(k+1)}=A\cdot y^{(k)} \end{cases} \quad (k=0,\ 1,\ 2,\ \cdots) \qquad (10.3.19)$$

容易验证：$y^{(k)}=A^k\cdot x^{(0)}/\max(A^k\cdot x^{(0)})$。所以，由前述分析得知：

$$y^{(k)}\sim x_1$$

同时，因为 $x^{(k+1)}=Ay^{(k)}\sim Ax_1=\lambda_1x_1\sim\lambda_1y^{(k)}$，所以，$x^{(k+1)}$ 的按模最大分量与 λ_1 相差一个可以忽略的小量，即

$$\max(x^{(k+1)})\sim\lambda_1$$

(2) 如果按模最大的特征值是一对共轭复根。此时，实初始向量 $x^{(0)}$ 的展式必可写为：

$$x^{(0)}=a_1x_1+\bar{a}_1\cdot\bar{x}_1+\sum_{i=3}^{n}a_ix_i$$

（注意，所讨论的是实矩阵情况。）仍按(10.3.19)式进行实运算，容易看出，此时

$$y^{(k)}\sim(a_1\lambda_1^kx_1+\bar{a}_1\bar{\lambda}_1^k\bar{x}_1)/\max(a_1\lambda_1^kx_1+\bar{a}_1\bar{\lambda}_1^k\bar{x}_1)$$

若令

$$\lambda_1=\rho e^{i\theta},\quad a_1=re^{i\alpha},\quad x_1=\{\xi_je^{i\varphi_j}\}$$

则得：

$$(a_1\lambda_1^kx_1+\bar{a}_1\bar{\lambda}_1^k\bar{x}_1)_j=2r\xi_j\rho^k\cos(\alpha+\varphi_j+k\theta)$$

由于 x_1 的各分量的幅角 φ_j 可能不同，所以，$y^{(k)}$ 的各分量在迭代过程中按先后顺序周期地变化其符号，迭代过程将不收敛。遇到这种情况，可按下述方法处理。

把 λ_1 与 $\bar{\lambda}_1$ 看作是某个二次方程：

$$\lambda^2-p\lambda-q=0$$

的根，其中 p, q 为实数。于是

$$(a_1\lambda_1^{k+2}x_1+\bar{a}_1\bar{\lambda}_1^{k+2}\bar{x}_1)-p(a_1\lambda_1^{k+1}x_1+\bar{a}_1\bar{\lambda}_1^{k+1}\bar{x}_1)-q(a_1\lambda_1^kx_1+\bar{a}_1\bar{\lambda}_1^k\bar{x}_1)$$
$$=a_1x_1(\lambda_1^{k+2}-p\lambda_1^{k+1}-q\lambda_1^k)+\bar{a}_1\bar{x}_1(\bar{\lambda}_1^{k+2}-p\bar{\lambda}_1^{k+1}-q\bar{\lambda}_1^k)=0$$

即是说，当 k 充分大时，$(\boldsymbol{A}^{k+2}-p\boldsymbol{A}^{k+1}-q\boldsymbol{A}^{k})\boldsymbol{x}^{(0)}\approx 0$ 再利用关系式

$$\max(\boldsymbol{A}^{k}\cdot\boldsymbol{x}^{(0)})=\max(\boldsymbol{x}^{k})\cdot\max(\boldsymbol{x}^{(k-1)})\cdots\max(\boldsymbol{x}^{(0)})$$

和

$$\boldsymbol{y}^{(k)}=\boldsymbol{A}^{k}\cdot\boldsymbol{x}^{(0)}/\max(\boldsymbol{A}^{k}\cdot\boldsymbol{x}^{(0)})$$

就可以得到

$$\max(\boldsymbol{x}^{(k+2)})\cdot\max(\boldsymbol{x}^{(k+1)})\cdot\boldsymbol{y}^{(k+2)}-p\cdot\max(\boldsymbol{x}^{(k+1)})\cdot\boldsymbol{y}^{(k+1)}-q\boldsymbol{y}^{(k)}\cong 0$$

这是两个未知数 p, q 的 n 个线性方程，从其中任选两个独立的方程求解或用最小二乘法求解，即可得出 p, q 之近似值。实际计算中，如果 k 相当大时，求得的 p, q 值稳定下来，就表明对应于 $\boldsymbol{x}_3$ 至 $\boldsymbol{x}_n$ 的分量已被消去，我们即可用下式求出 λ_1:

$$\mathrm{Re}(\lambda_1)=\frac{1}{2}p \qquad Im(\lambda_1)=\frac{1}{2}\sqrt{p^2+4q}$$

同时，可以验证如下向量 $\tilde{\boldsymbol{x}}_1$ 与特征向量 $\boldsymbol{x}_1$ 平行

$$\tilde{\boldsymbol{x}}_1=\frac{1}{2}\sqrt{p^2+4q}\cdot\boldsymbol{y}^{(k)}+i\left(\frac{1}{2}p\boldsymbol{y}^{(k)}-\boldsymbol{x}^{(k+1)}\right)$$

将其归一化就得到 $\boldsymbol{x}_1$。

最后我们要指出，上述方法的精确度往往是不理想的，特别是当 $Im(\lambda_1)$ 较小时精确度较差。因此，使用这一方法时要特别注意。

(3) 如果按模最大的特征值是实数，并且，其对应的初等因子是非线性的(此时特征向量不完全)，同时，其它特征值的模均严格地小于 $|\lambda_1|$，则幂法仍然收敛，但是收敛将很慢。这种情况下必须另找合适的算法。

(4) 幂法的加速与降阶

由(10.3.18)式可以看出，初始向量 $\boldsymbol{x}^{(0)}$ 在特征向量 $\boldsymbol{x}_j$ 上的分量按 $(\lambda_j/\lambda_1)^k$ 的速度收敛于零。因而，幂法是所谓“线性收敛”的方法，这类方法的收敛速度一般说来是不够理想的。特别是当 λ_2/λ_1 接近于 1 时，收敛将很慢，必须采取适当的加速措施。下面我们讨论两种常用的加速办法，为了简单起见，暂且假定:

$$\lambda_1>\lambda_2>\lambda_3\geqslant\lambda_4\geqslant\cdots\geqslant\lambda_n>0$$

经常使用的一种加速办法是所谓“δ^2 过程”。其基本思想是用相邻三次的近似向量进行组合来提高其精度，从而使过程得以加速。其办法如下：假设迭代过程已进行到 $\boldsymbol{x}_3\sim\boldsymbol{x}_n$ 上的分量可以忽略的程度，那么 $\boldsymbol{y}^{(k)}$ 将与 $\boldsymbol{x}_1+\varepsilon\boldsymbol{x}_2$ 相差一个常数因子。其中，ε 是一个小量；$\boldsymbol{x}_i$ 是按模最大元素为 1 的归一化向量。由于 ε 很小，于是有 $\max(\boldsymbol{x}_1+\varepsilon\boldsymbol{x}_2)=1+\varepsilon\cdot p$($p$ 为 $\boldsymbol{x}_2$ 的分量中与 $\boldsymbol{x}_1$ 按模最大分量相应者)；此外，以后的近似向量 $\boldsymbol{y}^{(k+l)}$ 中，按模最大分量的位置亦不再改变。这样，就可以把 $\boldsymbol{y}^{(k)}$、$\boldsymbol{y}^{(k+1)}$、$\boldsymbol{y}^{(k+2)}$ 写为:

$$(\boldsymbol{x}_1+\varepsilon\boldsymbol{x}_2)/(1+\varepsilon p),\ (\lambda_1\boldsymbol{x}_1+\varepsilon\lambda_2\boldsymbol{x}_2)/(\lambda_1+\varepsilon p\lambda_2),\ (\lambda_1^2\boldsymbol{x}_1+\varepsilon\lambda_2^2\boldsymbol{x}_2)/(\lambda_1^2+\varepsilon p\lambda_2^2)$$

现在，考虑如下向量 $\boldsymbol{z}$

$$\begin{cases}\boldsymbol{z}=(z_1,\ z_2,\ \cdots,\ z_n)^T\\ z_j=[y_j^{(k)}\cdot y_j^{(k+2)}-(y_j^{(k+1)})^2]/[y_j^{(k)}-2y_j^{(k+1)}+y_j^{(k+2)}]\\ \quad=y_j^{(k)}-[y_j^{(k)}-y_j^{(k+1)}]^2/[y_j^{(k)}-2y_j^{(k+1)}+y_j^{(k+2)}] \quad (j=1,\ 2,\ \cdots,\ n)\end{cases}$$

其中，$y_j^{(k)}$ 为向量 $y^{(k)}$ 的第 j 个分量。

把前面关于 $\boldsymbol{y}^{(k)}$、$\boldsymbol{y}^{(k+1)}$、$\boldsymbol{y}^{(k+2)}$ 的表达式代入上式，并以 $x_j^{(1)}$ 表示 $\boldsymbol{x}_1$ 的第 j 个分量，则可以得到

$$z_j=\left[x_j^{(1)}-\varepsilon^2\cdot\left(\frac{\lambda_2}{\lambda_1}\right)^2\cdot p\cdot x_j^{(2)}\right]\Big/\left[1-\varepsilon^2\cdot p^2\cdot\left(\frac{\lambda_2}{\lambda_1}\right)^2\right]$$

即是说，$\boldsymbol{z}=\boldsymbol{x}_1+O(\varepsilon^2)$。

因而，对于 $\boldsymbol{x}_1$ 来说，$\boldsymbol{z}$ 是较 $\boldsymbol{y}^{(k+2)}$ 更为精确的近似。这就是 δ^2 过程中的加速办法。

应该指出，上述过程是难于用程序控制其自动进行的，这是一个很大的缺点。不过，可以采用在计算机控制台上设立相应开关，由算题者视计算进行情况来决定是否采用上述加速过程的办法。有时，这一办法能取得较好的效果。

另外一个更为简单的加速办法是以 $\boldsymbol{A}-d\boldsymbol{I}$ 来代替矩阵 $\boldsymbol{A}$ 进行迭代。适当选取 d 也可使过程得以加速。很容易验证，此时有

$$(\boldsymbol{A}-d\boldsymbol{I})^k\cdot\boldsymbol{x}^{(0)}=(\lambda_1-d)^k\cdot\left(a_1\boldsymbol{x}_1+\left(\frac{\lambda_2-d}{\lambda_1-d}\right)^k a_2\boldsymbol{x}_2+\cdots+\left(\frac{\lambda_n-d}{\lambda_1-d}\right)^k a_n\boldsymbol{x}_n\right)$$

为了加速收敛，应使 $(\lambda_2-d)/(\lambda_1-d)$ 较 λ_2/λ_1 为小。我们只需令

$$d=\frac{1}{2}(\lambda_2+\lambda_n)$$

可以验证，此时

$$(\lambda_2-d)/(\lambda_1-d)=(\lambda_2-\lambda_n)/(2\lambda_1-\lambda_2-\lambda_n)<\lambda_2/\lambda_1$$

因而，过程得以加速。

这个办法也可用来求按模最小的特征值及相应特征向量，为此可令

$$d=\frac{1}{2}(\lambda_1+\lambda_{n-1})$$

此时，$(\boldsymbol{A}-d\boldsymbol{I})^k\cdot\boldsymbol{x}^{(0)}$ 的展式中最末项将居主导地位，过程就收敛于 $\boldsymbol{x}_n$。

上述加速办法称为移位法。由于特征值分布预先不知道，实际使用这个方法时会有困难，但因其简便，有时也使用。通常是使用者对特征值分布有一大概了解，以便粗略地估计一个 d 值。并且 d 值是通过计算机控制台上的手动开关给出的，每一次迭代开始时，程序从控制台读入一个 d 值。使用者可以根据迭代过程的进展随时修正所用 d 值（也可以由事先安排好的程序来修改），直到所用 d 值使迭代过程有明显加速为止。这种办法虽然有时是可以收到效果的，但由于需要手动开关配合，目前在计算机上已不大使用。

最后简单讨论一下求得 λ_1 以后如何进一步求出 λ_2, λ_3, $\cdots$ 的问题。这里只考虑实践中最常遇到的对称矩阵情况。假定已求得 λ_1 及 $\boldsymbol{x}_1$，则可构造如下矩阵：

$$\boldsymbol{A}^{(1)}=\boldsymbol{A}-\lambda_1\boldsymbol{x}_1\boldsymbol{x}_1^T/(\boldsymbol{x}_1^T\boldsymbol{x}_1)$$

因为对称矩阵的性质，故有 $\boldsymbol{x}_1^T\boldsymbol{x}_i=0\ (i\neq1)$ 因此，

$$\boldsymbol{A}^{(1)}\boldsymbol{x}_1=\boldsymbol{A}\boldsymbol{x}_1-\lambda_1\boldsymbol{x}_1(\boldsymbol{x}_1^T\cdot\boldsymbol{x}_1)/(\boldsymbol{x}_1^T\cdot\boldsymbol{x}_1)=0$$

$$\boldsymbol{A}^{(1)}\boldsymbol{x}_i=\boldsymbol{A}\boldsymbol{x}_i-\lambda_1\boldsymbol{x}_1(\boldsymbol{x}_1^T\cdot\boldsymbol{x}_i)/(\boldsymbol{x}_1^T\cdot\boldsymbol{x}_1)=\lambda_i\boldsymbol{x}_i\quad(i\neq1)$$

这样，矩阵 $\boldsymbol{A}^{(1)}$ 的按模最大特征值变为 λ_2，以 $\boldsymbol{A}^{(1)}$ 代替 $\boldsymbol{A}$ 进行迭代将求得 λ_2 及 $\boldsymbol{x}_2$，如此等等。为了保留原始矩阵 $\boldsymbol{A}$ 的特性，不需将 $\boldsymbol{A}^{(1)}$ 明显求出来，只要直接作 $\boldsymbol{x}_1^T$ 与迭代向量的乘积即可，这对于高阶的稀疏矩阵有重要意义。应当指出，用这种方法求出的 λ_2、$\boldsymbol{x}_2$ 精度一般已较 λ_1、$\boldsymbol{x}_1$ 差，若继续使用此法求 λ_3、$\boldsymbol{x}_3$ 等，精度将更差。因此，只能使用少数几次，以求出矩阵的前几个特征值和特征向量，而且，此法不如后面将要叙述的同时迭代法有效。

（二）反幂法

由于矩阵 $\boldsymbol{A}^{-1}$ 的特征值是 $1/\lambda_i$，所以 $\boldsymbol{A}^{-1}$ 的按模最大特征值为 $1/\lambda_n$。如果将 $\boldsymbol{A}$ 换为

A^{-1} 来进行幂法中的迭代，自然求得的特征值就是 $1/\lambda_n$，特征向量就是 x_n。也可以用移位法来加速迭代过程或者求其它的特征值，亦即用矩阵 $(A-dI)^{-1}$ 来进行迭代，这时求得的特征值为 $1/(\lambda_s-d)$，λ_s 为 λ_i 中与 d 最接近者。这就是反幂法的基本思想。

如上所述，反幂法的计算公式为：

$$\begin{cases} y^{(k)}=x^{(k)}/\max(x^{(k)}) \\ x^{(k+1)}=(A-dI)^{-1}\cdot y^{(k)} \end{cases} \quad (k=0,\ 1,\ 2,\ \cdots)$$

实际计算时，是用解方程组的办法来求 $x^{(k+1)}$ 的，即用如下公式：

$$\begin{cases} y^{(k)}=x^{(k)}/\max(x^{(k)}) \\ (A-dI)\cdot x^{(k+1)}=y^{(k)} \end{cases} \quad (k=0,\ 1,\ 2,\ \cdots) \tag{10.3.20}$$

为了节省工作量，常常先用列主元素消去法将矩阵 $(A-dI)$ 分解为下三角矩阵 L 与上三角矩阵 U 的乘积：

$$P\cdot(A-dI)=L\cdot U$$

其中矩阵 P 为形如(10.3.3)的一系列变换矩阵的乘积。这样，在迭代过程的每一步，只需解如下两个三角方程组：

$$\begin{cases} Lz=P\cdot y^{(k)} \\ Ux^{(k+1)}=z \end{cases}$$

反幂法经常用来求高阶稀疏矩阵的最小特征值及特征向量。在这种场合，系数矩阵通常是不存放起来的，事先的分解也就往往不可能，而只能在迭代过程的每一步，解一次方程组。由于此时系数矩阵总是具有某种特殊形状，在解方程组时需选取适当方法。关于这个问题，可以参看本书§8.1。

反幂法还经常用于知道某个特征值 λ_i 的近似值去求其相应的特征向量。现简单讨论一下这个过程的某些特点。假设已知 λ_1 的近似值 μ，任意选取一个初始向量 $x^{(0)}$，并令 $d=\mu$，按(10.3.20)式进行迭代。首先遇到的问题是：迭代过程的每一步，必须求解一个系数矩阵近于奇异的线性方程组。这是由于 $\mu\approx\lambda_1$，故 $\det(A-\mu I)\approx 0$，因而，在解方程组时我们必须采取某些措施。例如，采用主元素消去法以及防止由于过小的主元素而产生上溢等等。但无论如何求解一个系数矩阵近于奇异的方程组总是困难的，一般说来，所得解答中误差总占优势。好在此时解答中误差的主导方向也恰好是特征向量 x_1 的方向，而这正好是我们所要求的最终结果。所以，解方程组的不准确不影响迭代过程的进行。其次，再来看一看过程的收敛情况。为简单起见，假定矩阵 A 的初等因子都是线性的，也即是说，

$$x^{(0)}=\sum_{i=1}^{n}a_i x_i。$$

从(10.3.20)很容易得到第一次迭代的结果

$$x^{(1)}=c^{(1)}\left(a_1x_1+(\lambda_1-\mu)\cdot\sum_{i=2}^{n}\frac{a_ix_i}{\lambda_i-\mu}\right)$$

显然，当 $\lambda_1-\mu$ 很小且 a_1 不是小量时，括号中第二项相对于第一项已是小量，并且 $\lambda_1-\mu$ 愈小第二项愈小。通常若机器的字长为 t，则 $(\lambda_1-\mu)/\mu\approx 2^{-t}$，所以第二项与第一项的比值已接近于 2^{-t} 量级，即只迭代一次就达到相当高的精度，实际计算的大多数情况都是少数几次迭代就达到要求的精度，反幂法在这种场合是十分有效的。如果相应于 λ_1 的初等因子是非线性的或者 λ_1 是一个病态的特征值(例如与其它 λ_i 很靠近等等)，往往是第一次迭代改进较

大，其后收敛将较慢一些，但多数情况下仍能较快求得结果。所以，知道特征值的近似值去求相应的特征向量时，反幂法是一个比较有效的方法。

最后指出，将幂法与反幂法结合使用有时能够取得很好的效果。其具体作法是先选择一个适当的 d 值用幂法迭代，求得特征值的近似值 μ 及相应的特征向量 $\boldsymbol{x}$，然后，令 $d=\mu$ 和 $\boldsymbol{x}^{(0)}=\boldsymbol{x}$，进行反幂法的迭代，这样就能较快地求得一个准确的特征向量及特征值。

(三)多向量的同时迭代法

如果要用幂法求出矩阵 $\boldsymbol{A}$ 的前 p 个特征值与特征向量，则有时可以采用降阶的办法。但在降阶过程中，为了保持原来矩阵的某些特点，通常存储器中需留出 $n\times p$ 个单元来保存前面求得的特征向量。同时，往往预先不知道当时要求的特征值是实的、复的或其相应的初等因子是否非线性等等，所以，整个过程较难用程序控制自动进行。此外，后面几个特征向量的精度受到前面计算的影响也往往较差。由于存在这些缺点，人们自然想到是否可用 p 个初始向量同时进行幂法中的迭代来直接求出前 p 个特征值及特征向量⊖。由于这样作时，所需存储量仍为 $n\times p$，而计算过程的每一步都与原始矩阵 $\boldsymbol{A}$ 直接发生联系，有利于提高结果精度，又由于 p 个向量一起进行迭代，能够提供更多的信息来判断特征值的分布等等，一般说来，会比降阶过程更为可取一些。这种方法就是通常所谓的“同时迭代”。

显然，为了在迭代过程中获得尽可能多的信息，p 个初始向量应选成线性无关的。并且，每迭代一步后，希望仍旧保持其为线性无关。可以采用 p 个向量进行线性组合或进行正交化的办法来作到这一点。在一定条件下，可以证明它们将收敛于矩阵 $\boldsymbol{A}$ 的前 p 个特征向量。

如上所述，很容易写出两种同时迭代法的计算公式。

(1) 梯形化法

假设 $\boldsymbol{X}_0$ 为初始向量列构成的 $n\times p$ 矩阵，为保证其列为线性无关，则可将其取为如下“梯形”形状：

$$\boldsymbol{X}_0=\begin{bmatrix} \overbrace{1 & & & }^{p} \\ \times & 1 & & \mathbf{0} \\ \times & \times & 1 & \\ \times & \times & \times & \ddots \\ \vdots & \vdots & \vdots & 1 \\ & & & \times \\ & & & \times \\ & & & \vdots \\ \times & \times & \times\cdots\cdots & \times \end{bmatrix}$$

迭代一步后得到：$\widetilde{\boldsymbol{X}}_1=\boldsymbol{A}\boldsymbol{X}_0$。为保证 $\widetilde{\boldsymbol{X}}_1$ 的各列线性无关，可对其列进行线性组合(即对其列施行高斯消去过程)，再将其化为上面的“梯形”。这相当于对 $\widetilde{\boldsymbol{X}}_1$ 右乘以某个 $p\times p$ 的上三角矩阵 $\boldsymbol{U}_1^{-1}$，即有

$$\begin{cases} \boldsymbol{A}\boldsymbol{X}_0=\widetilde{\boldsymbol{X}}_1 \\ \boldsymbol{X}_1=\widetilde{\boldsymbol{X}}_1\cdot\boldsymbol{U}_1^{-1} \end{cases}$$

⊖ 实际计算时，为求出 p 个特征值，往往选取多于 p 个的向量同时进行迭代。

或写为 $$AX_0=X_1\cdot U_1$$

于是，一般的迭代公式可以写为

$$AX_k=X_{k+1}\cdot U_{k+1}\quad (k=0,\ 1,\ 2,\ \cdots)\tag{10.3.21}$$

这一方法通常称为梯形幂法。

如果要求出全部特征值和特征向量，自然应令 $p=n$，初始矩阵 X_0 及其后的逐次迭代矩阵 X_k 将为单位下三角矩阵，而 U_k 将为上三角形矩阵。这就是通常所谓的三角幂法。

(2) 正交化法

假设 Y_0 为初始向量列所构成的 $n\times p$ 矩阵，取其各列为相互正交的单位长度的向量。迭代一步后得到：$AY_0=\widetilde{Y}_1$。一般来说，$\widetilde{Y}_1$ 之各列不再正交。为了保证其各列线性无关，将 $\widetilde{Y}_1$ 的各列正交归一化，例如，用所谓的格拉姆-施密特(Gramm-Schmidt)正交化过程(参见§8.3)。这相当于对其右乘某个上三角矩阵 R_1^{-1}，即有

$$\begin{cases}AY_0=\widetilde{Y}_1\\ Y_1=\widetilde{Y}_1\cdot R_1^{-1}\end{cases}\quad \text{或写为}\quad AY_0=Y_1\cdot R_1$$

于是一般的迭代公式可以写为：

$$\begin{cases}AY_k=Z_{k+1}\\ Y_{k+1}=Z_{k+1}\cdot R_{k+1}^{-1}\end{cases}\quad \text{或}\quad AY_k=Y_{k+1}R_{k+1}\quad (k=0,\ 1,\ \cdots)\tag{10.3.22}$$

如果要求出全部特征值和特征向量，自然应令 $p=n$，初始矩阵 Y_0 及其后的逐次迭代矩阵 Y_k 将为 $n\times n$ 的正交矩阵，R_k 为 $n\times n$ 上三角形矩阵。这就是通常所谓的正交幂法。

实际使用同时迭代法的经验表明，对于一般矩阵，此法虽有效果，但对于对称正定矩阵，此法效果比较突出。特别是对于对称正定的阶数较高的稀疏矩阵，仅需求出其少量的（例如，远少于 $n/10$ 个）特征值和特征向量时，更是如此。在处理对称矩阵时（处理特征值问题时，正定的限制是无关紧要的）采用下列步骤的正交化方法往往比式(10.3.22)更为有效：

(i) 计算 $Z_{k+1}=AY_k$

(ii) 求出 $p\times p$ 对称正定矩阵 $G_{k+1}=Z_{k+1}^T\cdot Z_{k+1}$。并用旋转法(§10.3.4)。求出其特征值和特征向量，即：

$$U_{k+1}^T\cdot G_{k+1}\cdot U_{k+1}=D_{k+1}^2$$

其中 D_{k+1} 为对角线矩阵，其对角线元的平方依次为 G_{k+1} 的由大至小排列的特征值，U_{k+1} 之相应列为 G_{k+1} 的特征向量。

(iii) 计算 $Y_{k+1}=Z_{k+1}\cdot U_{k+1}\cdot D_{k+1}^{-1}$

这样求得的 Y_{k+1}，显然是列正交归一的，这是因为

$$Y_{k+1}^T\cdot Y_{k+1}=D_{k+1}^{-1}\cdot U_{k+1}^T\cdot Z_{k+1}^TZ_{k+1}\cdot U_{k+1}\cdot D_{k+1}^{-1}=I_p$$

上述正交化的方法需要多作一次矩阵乘法和解一个 $p\times p$ 的对称特征值问题[即其中的(ii)]，其计算量自然是较大些。因而，往往采取先按式(10.3.22)迭代若干步，然后按上述方法迭代一步的办法交替进行之。可以证明(见[7]、[10])，若 $\lambda_p>\lambda_{p+1}$，按照这种方法进行迭代，Y_{k+1} 的各列最终将落入与矩阵 A 按模最大的前 p 个特征值 $\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_p$ 相对应的特征向量 $x_1,\ x_2,\cdots,\ x_p$ 所张的不变子空间内。特别是，若 $\lambda_1>\lambda_2>\cdots>\lambda_p>\lambda_{p+1}$，则 Y_k 的各列对应地收敛于 $x_1, x_2,\ \cdots,\ x_p$，R_k 将收敛于对角型矩阵，其(或 D_{k+1} 的)对角线元素依次收敛于 $\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_p$。可以用 Y_{k+1} 的各列是否均已落入 Y_k 之列所张的子空间来控制迭

代过程的结束。同时，还可以根据 $\boldsymbol{R}_{k+1}$ 收敛于对角型的情况来判定 $\boldsymbol{Y}_{k+1}$ 的哪些列已收敛至要求的特征向量。

同时迭代法也可以用来求按模最小的 p 个特征值和特征向量。只需将式(10.3.22)以及将第二种正交化方法的(i)分别换为按反幂法进行迭代之如下相应公式：

$$\begin{cases} \boldsymbol{A}\boldsymbol{Z}_{k+1}=\boldsymbol{Y}_k \\ \boldsymbol{Y}_{k+1}=\boldsymbol{Z}_{k+1}\cdot\boldsymbol{R}_{k+1}^{-1} \end{cases}$$

求解 $\boldsymbol{Z}_{k+1}$: $\boldsymbol{A}\cdot\boldsymbol{Z}_{k+1}=\boldsymbol{Y}_k \quad (k=0, 1, 2, \cdots)$

其它公式均无需改变。

(四)QR 方法

QR 方法是幂法的一种推广和变形。它可以用来求任意矩阵的全部特征值。这个方法的计算公式经过适当扩展，能够非常有效地解决阶数不太高的任意实矩阵的全部特征值问题，它也是目前解这类问题最有效的方法之一。下面对这一方法的基本思想及其实际运用问题作一初步介绍。许多细节可以参阅[6]的卷 II 以及[9、10]。此外，由于 LR 方法与 QR 方法有密切联系，故将这两个方法的有关部分一并讨论。

所谓 QR 方法(或 LR 方法)的基本计算步骤如下：先将矩阵 $\boldsymbol{A}=\boldsymbol{A}_1$ 分解为正交矩阵 $\boldsymbol{Q}_1$(或下三角阵 $\boldsymbol{L}_1$)与上三角矩阵 $\boldsymbol{R}_1$ 的乘积：$\boldsymbol{A}_1=\boldsymbol{Q}_1\boldsymbol{R}_1$(或 $\boldsymbol{A}_1=\boldsymbol{L}_1\cdot\boldsymbol{R}_1$)，然后将所得的因式矩阵 $\boldsymbol{Q}_1$(或 $\boldsymbol{L}_1$)与 $\boldsymbol{R}_1$ 逆序相乘，得出矩阵 $\boldsymbol{A}_2=\boldsymbol{R}_1\cdot\boldsymbol{Q}_1$(或 $\boldsymbol{A}_2=\boldsymbol{R}_1\cdot\boldsymbol{L}_1$)。这样就完成了 QR 方法(或 LR 方法)的一步计算。以 $\boldsymbol{A}_2$ 代替 $\boldsymbol{A}_1$，重复上述步骤即可得出 $\boldsymbol{A}_3$，以 $\boldsymbol{A}_3$ 代替 $\boldsymbol{A}_1$ 重复上述步骤即得出 $\boldsymbol{A}_4$，如此继续。在一定条件下可以证明这样得出的矩阵 $\boldsymbol{A}_k$ 将收敛于上三角形矩阵(注意，逐次的矩阵 $\boldsymbol{A}_k$ 间是相似的。例如，$\boldsymbol{A}_2=\boldsymbol{R}_1\boldsymbol{Q}_1=\boldsymbol{Q}_1^{-1}(\boldsymbol{Q}_1\boldsymbol{R}_1)\boldsymbol{Q}_1=\boldsymbol{Q}_1^{-1}\boldsymbol{A}_1\boldsymbol{Q}_1$)，其对角线元即为矩阵 $\boldsymbol{A}$ 的特征值。

如上所述，可以得知 QR(或 LR)方法的计算公式为：

$$\begin{cases} \boldsymbol{A}=\boldsymbol{A}_1=\boldsymbol{Q}_1\cdot\boldsymbol{R}_1 & (\text{或 } \boldsymbol{A}_1=\boldsymbol{L}_1\cdot\boldsymbol{R}_1) \\ \boldsymbol{R}_1\cdot\boldsymbol{Q}_1=\boldsymbol{A}_2=\boldsymbol{Q}_2\cdot\boldsymbol{R}_2 & (\text{或 } \boldsymbol{R}_1\cdot\boldsymbol{L}_1=\boldsymbol{A}_2=\boldsymbol{L}_2\cdot\boldsymbol{R}_2) \\ \quad\cdots\cdots & \quad\cdots\cdots \\ \boldsymbol{R}_k\cdot\boldsymbol{Q}_k=\boldsymbol{A}_{k+1}=\boldsymbol{Q}_{k+1}\cdot\boldsymbol{R}_{k+1} & (\text{或 } \boldsymbol{R}_k\cdot\boldsymbol{L}_k=\boldsymbol{A}_{k+1}=\boldsymbol{L}_{k+1}\cdot\boldsymbol{R}_{k+1}) \\ \quad\cdots\cdots & \quad\cdots\cdots \\ (k=1, 2, 3, \cdots) \end{cases} \tag{10.3.23}$$

前节讨论过的三角幂法和正交幂法，经过适当变形，就是上述的 LR 与 QR 方法。先来指出这一点。

在三角幂法中，每计算一步，需要作一个矩阵乘法 $\boldsymbol{A}\cdot\boldsymbol{X}_k$，然后将所得矩阵分解为单位下三角矩阵 $\boldsymbol{X}_{k+1}$ 与上三角矩阵 $\boldsymbol{U}_{k+1}$ 的乘积，下三角因子 $\boldsymbol{X}_{k+1}$ 即为新的近似值(见式(10.3.21))。这样，每次迭代需完成一个矩阵乘积和一次三角型分解。由于原始矩阵 $\boldsymbol{A}$ 必须保存，故所需存储量大约为 $2n^2$。但如果仅需求出特征值，上述格式是可以简化的。因为 $\boldsymbol{U}_k$ 的对角线元素将收敛至相应特征值，故 $\boldsymbol{U}_k$ 应该保存。$\boldsymbol{A}$ 及 $\boldsymbol{X}_k$ 则无需全部保存也可使计算进行下去。这一点从下述讨论可以得知。

将(10.3.21)改写为：

$$\begin{cases}(X_{k-1})^{-1}A\cdot X_{k-1}=(X_{k-1}^{-1}\cdot X_k)\cdot U_k \\ X_k^{-1}AX_k=(X_k^{-1}\cdot X_{k+1})\cdot U_{k+1} \qquad (X_k^{-1}\cdot X_{k+1})\text{为单位下三角矩阵} \\ \cdots\cdots\end{cases} \tag{10.3.24}$$

显然，等式左端的矩阵相似于 A，只需保留它们，就可求出 A 的特征值。因而，实际需要保存的矩阵是因式矩阵 $(X_k^{-1}\cdot X_{k+1})$ 和 U_{k+1}。再注意到若将第 k 步保存下来的矩阵 $(X_{k-1}^{-1}\cdot X_k)$ 和 U_k 逆序相乘之，即有

$$U_k\cdot(X_{k-1}^{-1}\cdot X_k)=X_k^{-1}\cdot A\cdot X_{k-1}\cdot(X_{k-1}^{-1}\cdot X_k)=X_k^{-1}AX_k$$

显然，这就是 $k+1$ 步应保存的矩阵 $(X_k^{-1}\cdot X_{k+1})$ 和 U_{k+1} 的乘积，将其进行三角形分解，就可求得 $(X_k^{-1}\cdot X_{k+1})$ 及 U_{k+1}。这样一来，若用 L_{k+1} 表示 $X_k^{-1}\cdot X_{k+1}$，用 R_{k+1} 表示 U_{k+1}，就可以把计算格式(10.3.24)改写为：

$$\begin{cases}\text{令 } X_0=I \\ X_0^{-1}AX_0=A_1\xrightarrow{\text{分解}}L_1\cdot R_1 \\ X_1^{-1}AX_1=R_1\cdot L_1=A_2\xrightarrow{\text{分解}}L_2\cdot R_2 \\ \cdots\cdots \\ X_{k-1}^{-1}AX_{k-1}=R_{k-1}\cdot L_{k-1}=A_k\xrightarrow{\text{分解}}L_k\cdot R_k \\ X_k^{-1}\cdot A\cdot X_k=R_k\cdot L_k=A_{k+1}\xrightarrow{\text{分解}}L_{k+1}\cdot R_{k+1} \\ \cdots\cdots\end{cases}$$

在一定条件下，L_k 将收敛于单位矩阵，R_k 的对角线元，因而，A_k 的对角线元素将收敛至相应的特征值。这就是式(10.3.23)中的 LR 方法。可以看出，每迭代一步的计算量是一次上三角矩阵 R_k 与下三角矩阵 L_k 的乘法和一次三角型分解，所需存储量约为 n^2。所以，无论运算量和存储量均较三角幂法有所节省。

完全类似地，可以将正交幂法中的矩阵 $Q_k=(Y_{k-1}^{-1}\cdot Y_k)$ 和 R_k 保存起来，并由它们逆序相乘后进行正交三角分解来得出 $Q_{k+1}=(Y_k^{-1}\cdot Y_{k+1})$ 和 R_{k+1}，这样就得出所谓 QR 方法的计算格式：

$$\begin{cases}\text{令 } Y_0=I \\ Y_0^TAY_0=A_1\xrightarrow{\text{分解}}Q_1\cdot R_1 \\ Y_1^TAY_1=R_1\cdot Q_1=A_2\xrightarrow{\text{分解}}Q_2\cdot R_2 \\ \cdots\cdots \\ Y_{k-1}^TAY_{k-1}=R_{k-1}\cdot Q_{k-1}=A_k\xrightarrow{\text{分解}}Q_k\cdot R_k \\ Y_k^TAY_k=R_k\cdot Q_k=A_{k+1}\longrightarrow Q_{k+1}\cdot R_{k+1} \\ \cdots\cdots\end{cases}$$

从上述讨论可以看到，LR 方法和 QR 方法实际上都是幂法的推广，特别是它们分别为三角幂法和正交幂法的变形。然而，在实际使用时，它们却更加有效些。

下面对这两个方法的基本性质和收敛性问题进行一些讨论，最后再对实际使用 QR 方法中的关键之处给以某些说明。

(1) 算法的基本性质

从前面关于 LR 及 QR 方法的叙述得知，这两个方法可用如下统一格式来建立。

将矩阵 $A_1=A$ 分解为因式矩阵的乘积 $F_1\cdot G_1$，其中 F_1 是非奇异的。那么，矩阵

$$A_2=G_1\cdot F_1=F_1^{-1}A_1F_1$$

将与 A_1 相似。对 A_2 重复此步骤，就可得出 A_3，如此继续，将得出如下矩阵序列，序列中任意矩阵均与 A_1 相似：

$$\begin{cases} A_1=A \\ A_k=F_k\cdot G_k \qquad (k=1,\ 2,\ \cdots) \\ A_{k+1}=G_k\cdot F_k \end{cases} \tag{10.3.25}$$

当我们采用第八章§8.1中的(LU)三角分解定理(定理1.1)来得出 F_k 及 G_k 时（此时 F_k 为单位下三角形矩阵，G_k 为上三角形矩阵)，就是所谓 LR 方法。采用(QR)正交三角分解定理1.2(此时 F_k 为正交矩阵，G_k 为有非负对角线元的上三角形矩阵)，则得出所谓 QR 方法。值得注意的是如果 A_1 是奇异矩阵，其秩为 r 且前 r 行线性无关，则 R_1 的后 $n-r$ 行应为零，故 QR 方法中矩阵 A_2 的后 $n-r$ 行必为零，其左上角 r 阶主子块将唯一确定。于是，可以仅对此子块施行 QR 方法。因而，正交三角分解定理中 A 为非奇的假定对于 QR 方法是无关紧要的。

矩阵序列(10.3.25)有如下两个基本性质：

(i)
$$\begin{aligned} A_{k+1}&=F_k^{-1}A_kF_k \\ &=F_k^{-1}F_{k-1}^{-1}A_{k-1}F_{k-1}F_k \\ &=\cdots\cdots \\ &=F_k^{-1}F_{k-1}^{-1}\cdots F_1^{-1}A_1F_1\cdots F_{k-1}F_k \end{aligned}$$

若令 $E_k=F_1F_2\cdots F_k$，则上式可写为：

$$A_{k+1}=E_k^{-1}A_1E_k$$

或者：

$$E_k\cdot A_{k+1}=A_1\cdot E_k \tag{10.3.26}$$

(ii) 若令 $H_k=G_k\cdot G_{k-1}\cdots G_1$，则有：

$$\begin{aligned} E_kH_k&=F_1F_2\cdots F_{k-1}F_kG_kG_{k-1}\cdots G_1 \\ &=E_{k-1}A_kH_{k-1} \\ &=A_1E_{k-1}H_{k-1} \quad (\text{利用性质(i)}) \\ &=A_1^2E_{k-2}H_{k-2} \\ &=\cdots\cdots \\ &=A_1^k \end{aligned} \tag{10.3.27}$$

显然，无论在 LR 或 QR 方法中，E_k 与 F_k，H_k 与 G_k 均有相同的形状与特性，从分解式的唯一性得知，E_k 与 H_k 必定就是 A_1^k 的相应分解因子。

从上述两个基本性质我们可以看到，序列中任意矩阵 A_{k+1} 与 A_1 间的相似变换矩阵 E_k，就是 A_1^k 的相应分解式中第一个因子矩阵。自然，只要 A_1 是实矩阵，A_k 亦应为实矩阵。用上述形式将 LR 与 QR 方法统一之后，我们即可对这两个方法的收敛性问题进行统一的讨论。

(2) 收敛性问题

我们现在讨论 LR 与 QR 方法的收敛性问题。显然，为了求得特征值，只要求序列$\{A_k\}$收敛于一种简单形式的矩阵，例如三角型(或分块三角型)，而其对角线元(或块)有确定极限

即可。因而，我们这里约定，只要 $\{A_k\}$ 收敛于三角型(或分块三角型)其对角线元(或子块)有确定极限，无论其对角线外元素(或子块)是否有确定极限，都叫作方法是收敛的(亦称为按型收敛或本质收敛)。

通过简单验算并考虑到分解式是矩阵元素的连续函数这一事实，可以得出如下引理：

引理 3.1：若 $A_1=I$ 为单位矩阵，则 LR 与 QR 方法中的因式矩阵 L_k, R_k; Q_k, R_k 均应为单位矩阵。此外，若序列 $\{A_k\}$ 收敛于单位矩阵，则其相应因子 L_k, R_k, Q_k, R_k 均收敛于单位矩阵。

我们先考虑矩阵 A_1 的特征值按模不相等的情形。如下定理说明了 LR 与 QR 方法在这种情况下的收敛性质。

定理 3.2：假定：

① $A_1=A=X\Lambda X^{-1}$, $\Lambda=\mathrm{diag}\ (\lambda_1, \lambda_2, \cdots, \lambda_n)$;

② $|\lambda_1|>|\lambda_2|>\cdots>|\lambda_n|>0$;

③ $Y=X^{-1}$ 有三角分解式：$Y=L_y\cdot U_y$(L_y 为单位下三角型，U_y 为上三角型矩阵)；
则 QR 方法是收敛的(本质收敛)。

此外，若再假定：

④ X 有三角分解式：$X=L_x\cdot U_x$(L_x 为单位下三角型，U_x 为上三角型矩阵)；
则 LR 方法也是收敛的。

证明：由于 $A_{k+1}=E_k^{-1}A_1E_k$，故矩阵 E_k 将决定序列$\{A_k\}$的收敛性质。矩阵 E_k 为 A_1^k 的分解式第一个因子，我们可以通过 A_1^k 的表达式求出它，进而分析序列$\{A_k\}$的收敛性质。

$$\begin{aligned}A_1^k&=X\Lambda^kX^{-1}\\&=X\Lambda^kL_yU_y\quad(\text{假定③})\\&=X(\Lambda^kL_y\Lambda^{-k})(\Lambda^kU_y)\end{aligned}$$

若令 $\Lambda^kL_y\Lambda^{-k}=I+B_k$，上式可写为：

$$A_1^k=X(I+B_k)(\Lambda^kU_y)\tag{10.3.28}$$

显然，B_k 为对角线元等于零的下三角型矩阵，其元素

$$b_{ij}=l_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k\quad(i>j)$$

同时，由假定②知 $|\lambda_i/\lambda_j|<1\ (i>j)$，故

$$\lim_{k\to\infty}B_k=0$$

我们先来证明 QR 方法的收敛性。应用第八章§8.1的正交三角分解定理 1.2，X 总可唯一地分解为：

$$X=Q_x\cdot R_x\quad(R_x\text{ 之对角线元为正})$$

所以

$$\begin{aligned}A_1^k&=Q_xR_x(I+B_k)(\Lambda^kU_y)\\&=Q_x(I+R_xB_xR_x^{-1})(R_x\Lambda^kU_y)\end{aligned}$$

当 k 充分大时，$I+R_xB_kR_x^{-1}$ 将为非奇的，所以它应有唯一的分解式：$\tilde{Q}_k\cdot\tilde{R}_k$，并且，由引理 3.1 得知：

$$\lim_{k\to\infty}\tilde{Q}_k=\lim_{k\to\infty}\tilde{R}_k=I$$

于是 $$A_1^k=(Q_x\tilde{Q}_k)(\tilde{R}_kR_x\Lambda^kU_y)\quad (k \text{ 充分大})$$

显然，可以找出对角型矩阵 D_1，D_2，使得 $D_1\Lambda$ 和 D_2U_y 之对角线元均为正数(例如可令 $D_1=\mathrm{diag}(e^{i\theta_1}, e^{i\theta_2}, \cdots, e^{i\theta_n})$，其中 $\theta_i=-\arg(\lambda_i)$)。注意到 $\tilde{R}_kR_x\Lambda^kU_y$ 为上三角型矩阵，因而，矩阵 $D_1^kD_2\tilde{R}_kR_x\Lambda^kU_y$ 的对角线元亦为正数。同时，我们有：

$$A_1^k=(Q_x\tilde{Q}_kD_2^{-1}D_1^{-k})(D_1^kD_2\tilde{R}_kR_x\Lambda^kU_y)$$

由于 D_1，D_2 为对角型酉矩阵，所以上式中第一个因子亦应为酉矩阵。根据正交三角分解定理 3.1 的唯一性论述，我们得知 $Q_x\tilde{Q}_kD_2^{-1}D_1^{-k}$ 应为 A_1^k 的正交因子。这样一来，便有：

$$A_{k+1}=D_1^kD_2Q_k^*Q_x^*\cdot A_1\cdot Q_x\tilde{Q}_kD_2^{-1}D_1^{-k}$$

或者

$$\lim_{k\to\infty}A_{k+1}=\lim_{k\to\infty}D_1^k\cdot(D_2R_x\Lambda R_x^{-1}D_2^{-1})\lim_{k\to\infty}D_1^{-k}$$

显然，上式括号中的矩阵为一与 k 无关的确定上三角型矩阵，左乘 D_1^k 和右乘 D_1^{-k} 后其对角线元素不变。因而，不管 D_1^k 有无确定极限，我们都能得知 A_{k+1} 本质地(或按型)收敛。这就是要证明的结果。

对于 LR 方法，情形更为简单。由假定 ④ 及(10.3.28)式，我们有：

$$A_1^k=L_xU_x(I+B_k)(\Lambda^kU_y)=L_x(I+U_xB_kU_x^{-1})(U_x\Lambda^kU_y)$$

当 k 充分大时，$I+U_xB_kU_x^{-1}$ 亦必有三角分解式 $\tilde{L}_k\tilde{U}_k$，并且，从引理 3.1 知

$$\lim_{k\to\infty}\tilde{L}_k=\lim_{k\to\infty}\tilde{U}_k=I$$

于是

$$A_1^k=(L_x\tilde{L}_k)(\tilde{U}_kU_x\Lambda^kU_y)\quad (k \text{ 充分大})$$

这就是 A_1^k 的一种三角分解式。同样，从唯一性论述我们可以得知($L_x\tilde{L}_k$)为 A_1^k 分解式中第一个下三角因子。这样一来，便有：

$$A_{k+1}=\tilde{L}_k^{-1}L_x^{-1}A_1L_x\tilde{L}_k=\tilde{L}_k^{-1}(U_x\Lambda U_x^{-1})\tilde{L}_k$$

即是说，

$$\lim_{k\to\infty}A_k=U_x\Lambda U_x^{-1}$$

这就是要证明的结果。

关于上述两个方法在更一般情形下的收敛性问题，我们不再详细讨论，感兴趣的读者可以参阅[3]，[6]的卷 II，[9]，[11]。我们这里仅叙述一下与 QR 算法有关的一些结论：

① 对于任意方阵 A，QR 方法所产生的矩阵序列 $\{A_k\}$ 将按型收敛于分块上三角型矩阵，其对角线上每一子块有等模的特征值。

② 如果矩阵 A 等模的各特征值中，只有实的重特征值或复共轭的重特征值对，则上述的对角线子块将收敛于上三角型或 2×2 的分块上三角型。

③ 若矩阵 A 有若干组等模但互不相等的特征值，一般情况下，上述的对角线上子块将不收敛于上三角型或 2×2 分块上三角型。为求出特征值，此时应采取适当措施(见后面的讨论)。

最后，我们还要说明，上述定理的假定①、②成立时，通过对 B_k 及 $\tilde{Q}_k$ 等矩阵元素的估算，可以得到：

$$\begin{cases} a_{ii}^{(k)}=\lambda_i+O(r_i^k) \\ a_{i+1,i}^{(k)}=O(r_i^k) \end{cases}\quad (i=1, 2, \cdots, n)\qquad (10.3.29)$$

其中

$$r_i=\max\left\{\left|\frac{\lambda_i}{\lambda_{i-1}}\right|,\ \left|\frac{\lambda_{i+1}}{\lambda_i}\right|\right\}\quad(\lambda_0=\infty,\ \lambda_{n+1}=0)$$

即是说，$\boldsymbol{A}_k$ 的对角线元素将线性地收敛于矩阵 $\boldsymbol{A}$ 的特征值。

(3) 实际计算中所采取的措施

从现在起，我们只讨论 QR 方法，因为在实际计算中它是最常用的。对于 LR 方法，下面的许多讨论仍然适用，但这里我们不去涉及。

由前面讨论得知，QR 方法实际上是幂法的推广，其收敛速度是线性的。实际计算中，线性收敛速度很不理想，除非比值 r_i 很小。除此而外，每计算一步(即由 $\boldsymbol{A}_k\to\boldsymbol{A}_{k+1}$)，QR 方法的工作量亦很大，如果用镜像映射矩阵来实现分解(即按第八章 §8.1 的(8.1.30)式)，我们有 $\boldsymbol{H}_{n-1}\cdots\boldsymbol{H}_1\boldsymbol{A}_k=\boldsymbol{R}_k$，因而，$\boldsymbol{A}_{k+1}=\boldsymbol{R}_k\cdot\boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_{n-1}$，其所需乘法量大约为 $\frac{4}{3}n^3$。如果需要计算 n 步，则需完成 n^4 数量级的乘法运算，这是一个比较大的数字。所以前述的 QR 方法不仅收敛较慢，运算量也很大，若不采取适当措施，其实用价值是不大的。不过，上述的两个缺点，目前都有了较好的克服措施，采取这些措施后，QR 方法即成为实际计算中很有效的一种方法。通常把前面讨论的 QR 方法称之为基本 QR 方法，而采取下面各种措施以后的 QR 方法称之为扩展的 QR 方法。现在，我们分别叙述这些措施：

(i) 化为上海森堡型

QR 方法(同样 LR 方法)有一个重要特性，即是若 $\boldsymbol{A}_k$ 为上海森堡型矩阵，则 $\boldsymbol{A}_{k+1}$ 亦必为上海森堡型。这一事实很容易从(10.3.23)式看出，因为 $\boldsymbol{A}_k\boldsymbol{R}_k^{-1}=\boldsymbol{Q}_k$，$\boldsymbol{R}_k^{-1}$ 亦为上三角型矩阵，$\boldsymbol{Q}_k$ 的第 j 列必为 $\boldsymbol{A}_k$ 的前 j 列的线性组合，当 $\boldsymbol{A}_k$ 为上海森堡型时，其前 j 列的线性组合之形状必与其第 j 列相同，这样，$\boldsymbol{Q}_k$ 应与 $\boldsymbol{A}_k$ 有相同形状。再逆序相乘时，由于 $\boldsymbol{R}_k$ 为上三角型，将其右乘 $\boldsymbol{Q}_k$ 时，乘积矩阵的第 j 行将为 $\boldsymbol{Q}_k$ 的第 j 至第 n 行的线性组合，其形状仍与 $\boldsymbol{Q}_k$ 之第 j 行相同。这样，$\boldsymbol{A}_{k+1}$ 亦必与 $\boldsymbol{A}_k$ 有相同形状。

由于上述事实以及将基本 QR 算法用于上海森堡型矩阵时，每步计算量将仅为 n^2 数量级的乘法。所以，对于上海森堡型矩阵使用基本 QR 算法时，其每步运算量将自始至终为 n^2 的数量级。与对于一般矩阵的 n^3 级的乘法运算量相较，这是一个很大的节省。我们在 10.3.2 节已经看到，用镜像映射矩阵作相似变换很容易将一个矩阵化为上海森堡型，同时，理论分析与计算实践均证明这一变换是数值稳定的，其精确度很高。所以，在实际计算时，通常都先将矩阵化为上海森堡型，然后再使用 QR 方法求其特征值。为此，从现在起，我们将假定要讨论的矩阵 $\boldsymbol{A}$ 为上海森堡型矩阵。

(ii) 移位加速问题

为了提高基本 QR 方法的收敛速度，通常采用移位的办法。从前面讨论知道，如果我们对于 $\boldsymbol{A}_1-s\boldsymbol{I}$ 运用基本 QR 方法，那么，其第 n 行与 $n-1$ 列交叉处的元素将按 $(|\lambda_n-s|/|\lambda_{n-1}-s|)^k$ 的形式收敛于零。当 s 很接近 λ_n 时，这一元素自然会很快趋向于零，于是最右下角元素就是要求的一个特征值(注意，我们现在只考虑上海森堡型矩阵)。还可以对左上角的 $n-1$ 阶子矩阵进行同样处理，如此逐次降阶，最多 $n-1$ 次后即可求出所有特征值来。自然，其中每次的移位量 s 是无法预先确定的，但却可以从计算过程中逐步估计出来。即是我们不把移位量 s 取为常数，而是根据计算过程的进展，每迭代一步换一个数值，使其逐步逼近 λ_n。当然，还应保证每一步变换后的矩阵 $\boldsymbol{A}_{k+1}$ 与原来矩阵 $\boldsymbol{A}_1$ 相似。由于上述这些原

因，导致如下带恢复的移位加速格式：

$$\begin{cases} A_1 = A \\ A_k - s_k I = Q_k R_k \qquad (k=1,\ 2,\ \cdots) \\ A_{k+1} = R_k Q_k + s_k I \end{cases} \tag{10.3.30}$$

此时，很容易验证：

$$A_{k+1} = Q_k^{-1} Q_{k-1}^{-1} \cdots Q_1^{-1} \cdot A_1 \cdot Q_1 \cdots Q_{k-1} \cdot Q_k = E_k^{-1} A_1 E_k \tag{10.3.31}$$

$$E_k H_k = \prod_{i=1}^{k} (A_1 - s_i I) = \varphi_k(A_1)。其中\ \varphi_k(\lambda) = \prod_{i=1}^{k} (\lambda - s_i) \tag{10.3.32}$$

移位加速格式的收敛性也可按前节类似的办法来证明，不同之处仅在于用 $\varphi_k(A_1)$ 代替 A_1^k 来求出 E_k，然后再证明 $A_{k+1} = E_k^{-1} A_1 E_k$ 的收敛性，此处不再详述。对于移位加速格式，矩阵 A_k n 行 $(n-1)$ 列处的元素将按 $|\varphi_k(\lambda_n)/\varphi_k(\lambda_{n-1})|$ 的形式收敛。如果 s_i 很接近 λ_n，自然 $|\varphi_k(\lambda_n)|$ 很小，而分母非零，故收敛将得以加速。所以，对于上述带恢复的移位加速格式来说，我们应取 s_i 尽可能地接近 λ_n。一般情况下，A_k 的最右下角 2 阶子矩阵

$$C_2 = \begin{pmatrix} a_{n-1,n-1}^{(k)}, & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)}, & a_{n,n}^{(k)} \end{pmatrix}$$

的特征值中按模较小的一个将收敛于 λ_n。所以，我们可以取其为移位量 s_i，以保证 s_i 与 λ_n 逐步接近。迭代过程一直进行到 $a_{n,n-1}^{(k)} \sim 0$ 或 $a_{n-1,n-2}^{(k)} \sim 0$ 为止。前一情况下，$a_{nn}^{(k)}$ 将为 λ_n 的近似值，我们可对 $n-1$ 阶左上角子块继续计算。后一情况下，矩阵 C_2 的两个特征值即为 λ_{n-1}，λ_n 的近似值，我们可将 A_k 的最后两行和两列去掉，再继续计算。

某些矩阵用上述办法求出的移位量总是零，这样就达不到加速的目的。因而，在按上述办法移位若干次（例如十次）后，若仍不收敛，就应换一种特殊的移位办法计算几步，再继续用原来的移位格式（详见[10]中程序的处理）。

(iii) 双步 QR 方法——避免复运算问题

绝大多数实际问题中，矩阵 A 是实矩阵。如果按带恢复的移位加速格式计算，只要移位量 s_i 是实数，所有 A_k 也应该是实矩阵，尽管矩阵 A_1 的某些 λ_k 可能是复共轭的，我们不需进行复运算也可求得它们。然而，用前述的移位量决定办法，s_i 却可能是复数（即 C_2 的特征值为复共轭的），为了使整个计算过程仅仅使用实运算，我们必须采取一些措施。

显然，如果 C_2 的复共轭特征值为 s，$\bar{s}$，并假定我们作如下两步计算：

$$\begin{cases} A_1 - sI = Q_1 \cdot R_1 \\ A_2 = R_1 \cdot Q_1 + sI \\ A_2 - \bar{s} I = Q_2 \cdot R_2 \\ A_3 = R_2 \cdot Q_2 + \bar{s} I \end{cases}$$

则应有：

$$E_2 = Q_1 Q_2 \quad 为酉矩阵$$

$$A_3 = E_2^* A_1 E_2$$

$$E_2 H_2 = (A_1 - sI)(A_1 - \bar{s} I)$$

最后一个式子说明 E_2 应为实矩阵，因为它是实矩阵 $(A_1 - sI) \cdot (A_1 - \bar{s} I)$ 的正交因子。所以，从第二式得知 A_3 亦必为实矩阵。这样一来，如果我们能找到一个方法从 A_1 直接算出 A_3，复运算就可避免。下述引理将解决直接从 A_1 计算 A_3 的问题：

引理 3.2: 若 $\boldsymbol{A}$ 为任意矩阵, $\boldsymbol{Q}$ 为酉矩阵, $\boldsymbol{H}$ 为上海森堡矩阵, 并且 $\boldsymbol{H}=\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$。那么, 只要知道 $\boldsymbol{Q}$ 的第一列及矩阵 $\boldsymbol{A}$, 并要求 $\boldsymbol{H}$ 的下次对角线元素 $h_{j+1,j}$ 为正实数, $\boldsymbol{Q}$ 的其它各列及整个矩阵 $\boldsymbol{H}$ 将唯一确定。

证明这个引理是容易的, 我们用 $\boldsymbol{q}_j$ 来表示 $\boldsymbol{Q}$ 的第 j 列, 由等式 $\boldsymbol{A}\boldsymbol{Q}=\boldsymbol{Q}\boldsymbol{H}$ 得知:

$$\boldsymbol{A}\boldsymbol{q}_j=\sum_{i=1}^{j+1}h_{ij}\boldsymbol{q}_i$$

或者

$$h_{j+1,j}\cdot\boldsymbol{q}_{j+1}=\boldsymbol{A}\boldsymbol{q}_j-\sum_{i=1}^{j}h_{ij}\boldsymbol{q}_i=\tilde{\boldsymbol{q}}_{j+1}$$

若要求 $h_{j+1,j}$ 为正正实数, 由上式即可定出 $h_{j+1,j}$:

$$h_{j+1,j}=\|\tilde{\boldsymbol{q}}_{j+1}\|_2$$

因而,

$$\boldsymbol{q}_{j+1}=\left(\frac{1}{h_{j+1,j}}\right)\tilde{\boldsymbol{q}}_{j+1}$$

即是说, 从 $\boldsymbol{Q}$ 的第 1 列至第 j 列以及 $\boldsymbol{H}$ 的第 j 列对角线以上元素, 可以决定 $h_{j+1,j}$ 及 $\boldsymbol{q}_{j+1}$。而 $h_{ij+1}(i\leqslant j+1)$ 亦可按下式从 $\boldsymbol{q}_1, \cdots, \boldsymbol{q}_{j+1}$ 定出来:

$$h_{ij+1}=\boldsymbol{q}_i^*\boldsymbol{A}\boldsymbol{q}_{j+1}\quad(i=1, 2, \cdots, j+1)$$

这样, 仅由 $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_j$ 及矩阵 $\boldsymbol{A}$, 我们就可唯一决定 $h_{j+1,j}$, $h_{i,j+1}(i=1, 2, \cdots, j+1)$ 及 $\boldsymbol{q}_{j+1}$。对于 $j=1, 2, \cdots, n-1$ 重复上述步骤即可唯一地决定矩阵 $\boldsymbol{H}$ 及 $\boldsymbol{q}_2, \cdots, \boldsymbol{q}_n$ (注意: $\boldsymbol{q}_{n+1}$ 应自动为零, 因为不出现 $h_{n+1,n}$ 的有关项)。

现在我们应用上述引理来导出双步 QR 算法的计算公式。即是说我们要找出一个仅用实运算而直接从 $\boldsymbol{A}_1$ 算出 $\boldsymbol{A}_3$ 的公式来。从关系式:

$$\boldsymbol{A}_3=\boldsymbol{E}_2^*\boldsymbol{A}_1\boldsymbol{E}_2$$

得知, 若 $\boldsymbol{A}_3$ 的下次对角线元素为正实数, 且 $\boldsymbol{E}_2$ 的第一列已知, 根据引理 3.2, $\boldsymbol{A}_1$ 将唯一决定 $\boldsymbol{A}_3$ 和 $\boldsymbol{E}_2$。所以, 我们只要找出任一正交矩阵 $\boldsymbol{Q}$, 其第一列与 $\boldsymbol{E}_2$ 之第一列相同, 并且 $\boldsymbol{Q}^*\boldsymbol{A}_1\boldsymbol{Q}=\boldsymbol{B}$ 为上海森堡型矩阵, 其下次对角线元亦为正数, 那么便有: $\boldsymbol{Q}=\boldsymbol{E}_2$, $\boldsymbol{B}=\boldsymbol{A}_3$。

$\boldsymbol{A}_3$ 的下次对角线元素为正实数这一事实, 可以从正交分解定理及 $\boldsymbol{A}_1$ 的下次对角线元素为正得出。直接验算可知, 若 $\boldsymbol{A}_1$ 的下次对角线元素为 $a_{i+1,i}$, 则 $\boldsymbol{A}_2$ 的相应元素为: $\frac{R_{i+1,i+1}}{R_{ii}}\cdot a_{i+1,i}$, 其中 R_{ii} 为 $\boldsymbol{A}_1-s\boldsymbol{I}$ 的三角形因子 $\boldsymbol{R}_1$ 的对角线元素。只要 s 不是 $\boldsymbol{A}_1$ 的特征值, $\boldsymbol{R}_1$ 将是非奇异的, 其对角线元素必为正实数, 所以, $\boldsymbol{A}_2$ 之下次对角线元素大于零。同理, $\boldsymbol{A}_3$ 亦有此性质。当 s 为 $\boldsymbol{A}_1$ 的特征值时, 由于它是从 $\boldsymbol{C}_2$ 算出来的, 此时 $\boldsymbol{A}_1$ 之相应行已变为三角型, 我们将采取降阶措施, 并求得一个或两个特征值, 然后重新对阶数较低的 $\tilde{\boldsymbol{A}}_1$ 进行处理。所以, 总可以认为 $\boldsymbol{A}_3$ 的下次对角线元大于零。

$\boldsymbol{E}_2$ 的第一列也很易求得, 因为 $\boldsymbol{E}_2$ 是矩阵: $(\boldsymbol{A}_1-s\boldsymbol{I})\cdot(\boldsymbol{A}_1-\bar{s}\boldsymbol{I})$ 正交三角分解式的正交因子, 所以, 其第一列就是 $(\boldsymbol{A}_1-s\boldsymbol{I})\cdot(\boldsymbol{A}_1-\bar{s}\boldsymbol{I})$ 的第一列归一化的结果。简单计算可得 $\boldsymbol{E}_2$ 之第一列为:

$$\tilde{\boldsymbol{q}}_1/\|\tilde{\boldsymbol{q}}_1\|_2 \text{ 其中: } \tilde{\boldsymbol{q}}_1=\begin{pmatrix} a_{11}^2+a_{21}\cdot a_{12}-(s+\bar{s})a_{11}+s\cdot\bar{s} \\ a_{21}\cdot a_{11}+a_{22}\cdot a_{21}-(s+\bar{s})a_{21} \\ a_{32}\cdot a_{21} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

根据引理 3.2 我们仅需找出一个正交矩阵 $\boldsymbol{Q}$, 其第一列为 $\tilde{\boldsymbol{q}}_1/\|\tilde{\boldsymbol{q}}_1\|_2$, 使 $\boldsymbol{B}=\boldsymbol{Q}^*\boldsymbol{A}_1\boldsymbol{Q}$ 为有正下次对角线元素的上海森堡矩阵，则矩阵 $\boldsymbol{B}$ 就是我们要计算的 $\boldsymbol{A}_3$。

矩阵 $\boldsymbol{Q}$ 可按如下方式建立:

首先找出一个镜像映射矩阵 $\boldsymbol{P}_1$, 使其第一列为 $\tilde{\boldsymbol{q}}_1/\|\tilde{\boldsymbol{q}}_1\|_2$, 即是:

$$\boldsymbol{P}_1\cdot\boldsymbol{e}_1=\tilde{\boldsymbol{q}}_1/\|\tilde{\boldsymbol{q}}_1\|_2$$

从第八章 §8.1 的(8.1.28)和(8.1.29)以及由于 $\tilde{\boldsymbol{q}}_1$ 仅有前三个分量非零，可以得知 $\boldsymbol{P}_1$ 及 $\boldsymbol{B}_1=\boldsymbol{P}_1^*\boldsymbol{A}_1\boldsymbol{P}_1$ 必分别有如下形式:

$$\boldsymbol{P}_1=\left(\begin{array}{ccc|ccc}
\times & \times & \times & & & \\
\times & \times & \times & & \mathbf{0} & \\
\times & \times & \times & & & \\
\hline
 & & & 1 & & \\
 & \mathbf{0} & & & \ddots & \\
 & & & & & 1
\end{array}\right)$$

$$\boldsymbol{B}_1=\boldsymbol{P}_1^*\boldsymbol{A}_1\boldsymbol{P}_1=\begin{pmatrix}
\times & \times & \times & \times & \cdots & \times \\
\times & \times & \times & \times & \cdots & \times \\
\times & \times & \times & \times & \cdots & \times \\
\times & \times & \times & \times & \cdots & \times \\
 & & & \times & \ddots & \vdots \\
 & \mathbf{0} & & & \ddots & \vdots \\
 & & & & \times & \times
\end{pmatrix}$$

然后，再用一个第一列为 $\boldsymbol{e}_1$ 的镜像映射矩阵 $\boldsymbol{P}_2$, 使 $\boldsymbol{B}_2=\boldsymbol{P}_2^*\boldsymbol{B}_1\boldsymbol{P}_2$ 的第一列化为上海森堡型。从本章 10.3.2 节的(10.3.13)得知这是可行的，并且，$\boldsymbol{P}_2$ 与 $\boldsymbol{B}_2$ 将分别有如下形式:

$$\boldsymbol{P}_2=\begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & \times & \times & \times & & \\
0 & \times & \times & \times & \mathbf{0} & \\
0 & \times & \times & \times & & \\
\vdots & & & 1 & & \\
0 & \mathbf{0} & & & \ddots & 1
\end{pmatrix}$$

$$\boldsymbol{B}_2=\boldsymbol{P}_2^*\boldsymbol{B}_1\boldsymbol{P}_2=\begin{pmatrix}
\times & \cdots & \cdots & \cdots & \cdots & \cdots & \times \\
b_{21} & \times & \times & \times & \cdots & \cdots & \times \\
0 & \times & \times & \times & & & \vdots \\
\vdots & \times & \times & \times & & & \vdots \\
\vdots & \times & \times & \times & \times & & \vdots \\
\vdots & & & & \times & \ddots & \vdots \\
0 & \cdots & \cdots & \cdots & \cdots & \times & \times
\end{pmatrix}$$

其中 $b_{21}>0$。同样，可以再用前两列为 $\boldsymbol{e}_1$、$\boldsymbol{e}_2$ 的镜像映射矩阵 $\boldsymbol{P}_3$, 使 $\boldsymbol{B}_2$ 的第二列为上海森堡型，依此类推，$n-1$ 步后我们有:

$$B=P_{n-1}^{*}P_{n-2}^{*}\cdots P_{1}^{*}A_{1}P_{1}\cdots P_{n-2}P_{n-1}$$

$$=\begin{pmatrix} \times & \times & \cdots & \cdots & \times \\ b_{21} & \times & & & \vdots \\ & b_{32} & \times & & \vdots \\ & \mathbf{0} & \ddots & \ddots & \vdots \\ & & & b_{nn-1} & \times \end{pmatrix}$$

其中 $b_{i+1i}>0$。同时，矩阵 $Q=P_1P_2\cdots P_{n-1}$ 的第一列与 P_1 之第一列相同，即为 $\tilde{q}_1/\|\tilde{q}_1\|_2$。这样，矩阵 B 就是经两步移位 QR 变换后的矩阵 A_3。

很明显，当 C_2 的特征根为两个实数时，我们也可按上述两步变换的格式进行计算。实际计算时正是这样，即不管 C_2 的特征值是实数或共轭复数，总是每次进行两步 QR 变换。这就是所谓的双步 QR 算法。

从上面的讨论可以看到，矩阵 P_i 的形状是很简单的，除去一个三阶对角子块外，其它部分均与单位矩阵相同。矩阵 B_i 亦有类似特点，其与上海森堡型不同之处仅仅在于三个元素。这些特点可以用来简化计算公式和程序(详细计算公式请参见本章后面的程序 8)。

(iv) 小元素的略去问题

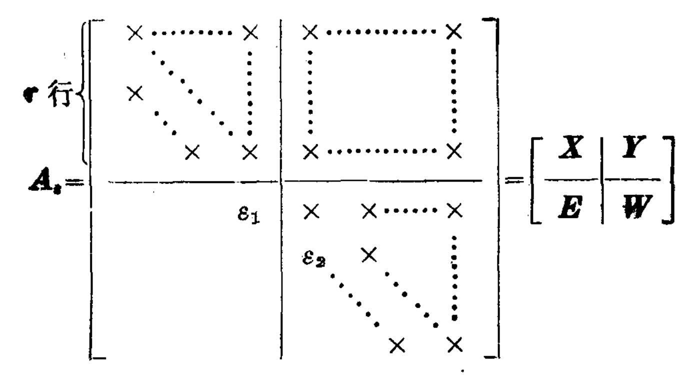
对于上海森堡型矩阵施行 QR 算法时，次对角线元素中的某些元素将随迭代过程的进展收敛于零。当这些元素充分小时，我们就可以将其忽略，而把问题化为低阶的情形。例如：

$$A_s=\left[\begin{array}{ccc|ccc|cccc} \times & \times & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \times \\ \times & \times & & & & & & & & \vdots \\ & \times & \times & & & & & & & \vdots \\ \hline & & \varepsilon & \times & & & & & & \vdots \\ & & & \times & \times & & & & & \vdots \\ & & & & \times & \times & & & & \vdots \\ \hline & & & & & \varepsilon & \times & & & \vdots \\ & & & & & & \times & \times & & \vdots \\ & & & & & & & \times & \times & \vdots \\ & & & & & & & & \times & \times \end{array}\right]\cong\begin{bmatrix} \mathrm{B} & \times & \times \\ 0 & \mathrm{C} & \times \\ 0 & 0 & \mathrm{D} \end{bmatrix}$$

由于 A_s 是由前面计算得来，其元素已有误差，当 ε 小于这个误差时，自然，以零代替它是无损于结果精度的。所以，一般取 $2^{-t}\cdot\|A\|$ 作为控制该元素应忽略与否的标准是合适的。

这样，在迭代过程的每一步，均去检查一遍下次对角线元素，若其最后一个可忽略的元素在位置 $(r+1,\ r)$ 处，则可对右下角 $n-r$ 阶子矩阵先进行计算，然后再处理左上角的 r 阶子矩阵。显然，$r=n-1$ 时，$a_{nn}^{(k)}$ 即为特征值，$r=n-2$ 时，C_2 的特征值即为矩阵 A_1 的特征值。

有时可能出现次对角线元素均不能忽略，但其中有某些相邻两个元素之乘积可以忽略的情况。例如，$(r+1,\ r)$ 及 $(r+2,\ r+1)$ 两元素之乘积可以忽略，矩阵 A_s 之形状如下：



可以证明其右下角子矩阵 $\boldsymbol{W}$ 的特征值 μ 也是某个矩阵 $\boldsymbol{A}'_s$ 的特征值，而矩阵 $\boldsymbol{A}'_s$ 与 $\boldsymbol{A}_s$ 仅在 $(r+2,\ r)$ 位置上相差一个元素 $\varepsilon_1\cdot\varepsilon_2/(a_{r+1,r+1}-\mu)$。显然，只要 $(a_{r+1,r+1}-\mu)$ 不是很小，此元素即可忽略。于是 μ 亦为 $\boldsymbol{A}_s$ 之近似特征值。这一事实之证明很简单，因为，

$$\boldsymbol{x}^T\cdot(\boldsymbol{W}-\mu\boldsymbol{I})=0\quad(\boldsymbol{x}\neq 0)$$

的第一个方程式为：

$$(a_{r+1,r+1}-\mu)x_1+\varepsilon_2x_2=0$$

或者写为：

$$\varepsilon_1x_1+\frac{\varepsilon_1\varepsilon_2}{a_{r+1,r+1}-\mu}x_2=0$$

这个方程与 $\boldsymbol{x}^T(\boldsymbol{W}-\mu\boldsymbol{I})=0$ 的 $n-r$ 个方程一起，说明 $\boldsymbol{A}'_s-\mu\boldsymbol{I}$ 的后 $n-r$ 行是线性相关的，即 μ 亦应为 $\boldsymbol{A}'_s$ 的特征值。利用上述事实，我们仍可在每迭代一步之后去检查一下相邻两元素乘积可忽略之条件是否满足，若满足条件，就把问题化为低阶的情形。

采取上述措施，除了节省工作量外，还可以提高某些情况下的收敛速度。因为次对角线元素很小，说明各子矩阵间关联很弱，如果不把它们分开来处理，则迭代过程中各子矩阵的特征值可能按统一次序逐个求出。这样，就可能出现要把前面某个子矩阵的特征值移至右下角来的情况，这时收敛会很慢。此外，次对角线元素非零，在前面讨论的双步 QR 算法中是必要的，将可忽略的元素除去后就能够保证这一点。

(v) 初始矩阵的平衡问题

舍入误差分析的结果表明，在不变矩阵的特征值的前提下，尽可能减少其范数是有利的。目前，大都用对行列引入比例因子(对角相似变换)的办法来减少矩阵的欧氏范数 $\|\boldsymbol{A}\|_E$。为了不引入额外的舍入误差，比例因子的数值应取为计算机上用以表示浮点数的基底之整幂次。这样引入的比例因子，将使矩阵达到近似地“平衡”，其欧氏范数也将有较大的减小(详见[10]pp. 315)。

上述各点即为通常使用 QR 方法时所采取的主要措施。由于这些措施，QR 方法已成为一个很有效的通用算法。许多实际计算的经验都说明，几乎对于所有的矩阵，QR 方法均收敛，并且，求出全部特征值的乘法量大约仅为 $C\cdot n^3$，其中 C 一般为小于 10 的常数。同时，计算结果也有很高的精确度。对于非对称的任意实矩阵，如阶数不高，整个矩阵可以放在计算机的内存中，使用 QR 方法求其全部特征值是很有效的。

最后，我们将 QR 方法的计算步骤作一概要的小结，以便读者了解整个算法：

(1) 用对角相似变换减少原始矩阵的欧氏范数，使矩阵达到近似的“平衡”，并保留变换矩阵(若需求特征向量)。

(2) 算出某些有用的常数,例如 $\|A\|_E$, trace(A), $2^{-t}\|A\|_E$ 等等。

(3) 用正交相似变换将矩阵 A 化为上海森堡型矩阵 A_1,并保留变换矩阵和矩阵 A_1(若需求特征向量)。

(4) 检查次对角线元素的某些是否已可忽略,并确定当前应处理的子矩阵的位置。

(5) 计算当前处理的子矩阵的最右下角二阶子块 C_2 的特征值。

(6) 判别是否已求得特征值。若已求得,则将其记录下来,并把要处理的子矩阵降阶,然后转去执行(4)。若已求得全部特征值,则转去执行(10),否则,继续作(7)。

(7) 选择位移量 s_{2k}, s_{2k+1}。并判别是否需作特殊位移处理。若必要,则作特殊处理。

(8) 执行一步双步 QR 变换(又称为迭代一次)。

(9) 迭代记数器加 1,并判别总迭代次数是否超过给定上限,若超过,则记录相应信息,作好求下一特征值的准备工作,并转(4);否则,直接转(4)。

(10) 印出求得的结果。例如,λ_i, $\Sigma\lambda_i-\text{trace}(A)$,迭代次数等等。

(11) 若需求特征向量,则用反幂法对上海森堡型矩阵 A_1 求出其特征向量,并用(1)和(3)中保留下来的变换矩阵将其还原为原始矩阵的特征向量。

我们给出按上述过程编制的仅求全部特征值的 QR 方法程序。由于只求特征值,所以,各次变换矩阵均不保留,只需 n^2 个单元存放原始矩阵,$3n$ 个单元分别存放特征值的实部虚部及迭代次数。程序的说明及程序本身请见本章最后所附的程序八。更详尽和完整的程序请见[10] pp. 372～395。

10.3.4 旋转法及其推广

旋转法是一种用平面旋转矩阵所构成的正交相似变换将对称矩阵化为对角型的方法,也叫作雅可比(Jacobi)法。由于其算法简单,行之有效,所以又将其推广至一般矩阵的特征值问题,即有所谓广义旋转法(或广义雅可比法)。本节主要讨论这两种方法。

(一)实对称矩阵的旋转法

(1) 我们知道,任意实对称矩阵,总可以通过正交相似变换化为对角线型。所谓旋转法本质上就是去设法找出这样一个正交矩阵 U, 使得 $U^TAU=\text{diag}(\lambda_i)$。于是,对称矩阵 A 的特征值就是 $\text{diag}(\lambda_i)$的对角元素,U 的各列就是相应的特征向量。根据这一想法,雅可比(1846 年)提出用一系列平面旋转矩阵的乘积来达到上述目的,下面,我们来说明其具体作法。

由于正交相似变换下,矩阵元素的平方和不变,即是说,若 $B=V^TAV$, V 为正交矩阵,则有:

$$N^2(A)=\sum_{i,j}a_{ij}^2=N^2(V^TAV)=N^2(B)=\sum_{i,j}b_{ij}^2$$

若能找出一种正交相似变换,使对称矩阵 A 在变换之后非对角线元素的平方和减少(自然,对角线元素平方和将增加),那么,因为正交相似变换保持矩阵的对称性,我们还可对变换后的矩阵继续施行上述变换,以进一步减小其非对角线元素的平方和。如此继续进行下去,最终将使非对角线元的平方和任意接近于零(对角线元平方和接近极大值),矩阵即变为近似的对角线型,这就是旋转法的基本思想。从(10.3.12)式可以看出,平面旋转矩阵能够达到这一目的。只要我们选择角度 φ,使得:

$$(a_{jj}-a_{ii})\sin 2\varphi+2a_{ij}\cos 2\varphi=0 \tag{10.3.33}$$

变换后矩阵非对角线元素平方和就将减少 $2a_{ij}^2$。所以，我们只需找出一个非零的 a_{ij}(自然，可以取 a_{ij} 为非对角线元中按模最大者，这样，非对角线元平方和减少最快)，并按如下公式确定角度 φ:

$$\operatorname{tg} 2\varphi=\frac{2a_{ij}}{a_{ii}-a_{jj}} \tag{10.3.34}$$

然后，用形如(10.3.7)的平面旋转矩阵 $\boldsymbol{J}_{ij}(\varphi)$ 作如下正交相似变换:

$$\boldsymbol{A}^{(1)}=\boldsymbol{J}_{ij}(\varphi)^T\cdot\boldsymbol{A}\cdot\boldsymbol{J}_{ij}(\varphi)$$

则 $\boldsymbol{A}^{(1)}$ 的非对角线元平方和 $S(\boldsymbol{A}^{(1)})$ 将为:

$$S(\boldsymbol{A}^{(1)})=\sum_{k\neq l}a_{kl}^{(1)2}=S(\boldsymbol{A})-2a_{ij}^2$$

再对 $\boldsymbol{A}^{(1)}$ 重复上述变换，得到对称矩阵 $\boldsymbol{A}^{(2)}$，其非对角线元平方和将比 $\boldsymbol{A}^{(1)}$ 的更小，等等。如此不断重复下去，我们就得到一个矩阵序列 $\{\boldsymbol{A}^{(k)}\}$。只要适当地限制 $a_{i_kj_k}^{(k-1)}$ 的取法，例如，取其为非对角线元中按模最大者，此时，由于:

$$S(\boldsymbol{A}^{(k)})=S(\boldsymbol{A}^{(k-1)})-2a_{i_kj_k}^{(k-1)2}\leqslant\left(1-\frac{2}{n(n-1)}\right)S(\boldsymbol{A}^{(k-1)})\leqslant\cdots$$

$$\leqslant\left(1-\frac{2}{n(n-1)}\right)^kS(\boldsymbol{A})$$

当 k 充分大时，$S(\boldsymbol{A}^{(k)})$ 之值便可以任意小。因而，上述过程重复足够多次后，矩阵 $\boldsymbol{A}^{(k)}$ 就将近似于对角型。

由上述讨论得知，旋转法的计算步骤可以归结为:

(i) 在矩阵 $\boldsymbol{A}$ 中找出一个非零的非对角线元素 a_{ij}，例如，取 a_{ij} 为按模最大的非对角线元素(并由其所在行、列定出下标 i, j)。

(ii) 由条件:

$$(a_{jj}-a_{ii})\cdot\sin 2\varphi+2a_{ij}\cdot\cos 2\varphi=0$$

定出 $\sin\varphi$ 与 $\cos\varphi$。

(iii) 按如下公式计算矩阵 $\boldsymbol{A}^{(1)}$ 的元素 $a_{kl}^{(1)}$:

$$\boldsymbol{A}^{(1)}=\boldsymbol{J}_{ij}^T(\varphi)\cdot\boldsymbol{A}\cdot\boldsymbol{J}_{ij}(\varphi)$$

$$\begin{cases}a_{kl}^{(1)}=a_{kl}\quad(k,\ l\neq i,\ j)\\ a_{il}^{(1)}=a_{il}\cos\varphi+a_{jl}\sin\varphi\\ a_{jl}^{(1)}=-a_{il}\sin\varphi+a_{jl}\cos\varphi\\ a_{li}^{(1)}=a_{li}\cos\varphi+a_{lj}\sin\varphi\\ a_{lj}^{(1)}=-a_{li}\sin\varphi+a_{lj}\cos\varphi\end{cases}\quad(l\neq i,\ j)$$

$$\begin{cases}a_{ii}^{(1)}=a_{ii}\cos^2\varphi+2a_{ij}\sin\varphi\cos\varphi+a_{jj}\sin^2\varphi\\ a_{jj}^{(1)}=a_{ii}\sin^2\varphi-2a_{ij}\sin\varphi\cos\varphi+a_{jj}\cos^2\varphi\\ a_{ij}^{(1)}=a_{ji}^{(1)}=0\end{cases} \tag{10.3.35}$$

(iv) 以 $\boldsymbol{A}^{(1)}$ 代替 $\boldsymbol{A}$，重复(i)(ii)(iii)求出矩阵 $\boldsymbol{A}^{(2)}$，如此等等。直到 $\max\limits_{p\neq q}|a_{pq}^{(k)}|\leqslant\varepsilon$ 时停止计算，此时，对角线元素即为要求的特征值。逐次变换矩阵 $\boldsymbol{J}^{(k)}$ 之乘积:

$$\boldsymbol{U}=\boldsymbol{J}^{(1)}\cdot\boldsymbol{J}^{(2)}\cdots\boldsymbol{J}^{(N)}$$

即为要求的特征向量。

可以看出，旋转法的计算步骤是很简单的。此外，经典的雅可比法(即取按模最大的非

对角线元素为 a_{ij} 者)，不论矩阵 $\boldsymbol{A}$ 的特征值分布如何，总是收敛的。不仅如此，还可以证明其具有所谓"渐近平方收敛速度"。即是说，进行足够多次变换后，若非对角线元素与

$$\rho=\min_{\lambda_i\neq\lambda_j}|\lambda_i-\lambda_j|$$

相较已为小量 ε，则最多再进行 $n(n-1)/2$ 次变换，非对角线元素均应变为 ε^2 级小量。因此，旋转法的收敛速度还是比较快的。更重要的是旋转法对舍入误差的影响有较强的稳定性，求得的结果精度一般都很高。特别是求得的特征向量正交性很好，这是其他方法所不及的。所以，旋转法是求对称矩阵全部特征值及特征向量的一个较好方法。旋转法也有其不足之处，首先是它不能有效利用矩阵的各种特殊形状，例如带状或稀疏性等，以节省工作量及提高能够处理的问题的阶数，因为在旋转法的计算过程中一般将破坏这些特性。所以，旋转法只能用于阶数不高的"满矩阵"。其次，是其计算工作量仍较大。它一般不如后面所述的化为三对角型的方法节省时间，特别是只需求特征值或只需求少量特征值和相应特征向量时更是如此。最后，绝对值较小的特征值精度差一些，虽然实际计算中采取了一些措施来提高精度，但很多情况下仍不理想。由于旋转法有上述的优缺点，我们在使用它时，就应根据问题的特点进行具体分析，以发挥其优点，克服其缺点。

(2) 实际计算中常常采用一些办法来节省工作量和提高精确度。常用的大致有下列两方面：

(i) 减少选取 $\max\limits_{k\neq l}|a_{kl}|$ 的机器时间。

由于每次变换前要找出 $\max\limits_{k\neq l}|a_{kl}|$ 的所在位置 i, j，需要完成许多运算，这是很耗费机器时间的。有几种办法可以克服这个缺点。一种是增加寄存单元的办法，即用 $2n$ 个单元来保存各列的按模最大元素及其所在行数。经过一次变换后仅有两列元素改变，故只需找出这两列的按模最大元素即可从 n 个单元中找到 $\max\limits_{k\neq l}|a_{kl}|$。这样做可以节省一些机器时间，但程序将要复杂化。另一种办法是不去找 $\max\limits_{k\neq l}|a_{kl}|$，而是按行(或列)的顺序依次取出非零的非对角线元素作为 a_{ij} 去进行相应的变换。这时足标对 (i, j) 将为(按行次序)：$(1, 2)$，$(1, 3)$，$\cdots$，$(1, n)$，$(2, 3)$，$(2, 4)$，$\cdots(2, n)\cdots(n-1, n)$。这种办法通常称之为循环的雅可比法。其优点自然在于完全省去了寻找 $\max\limits_{k\neq l}|a_{kl}|$ 的运算，但由此带来的缺点是不论 a_{ij} 的大小如何，均要进行相应的变换。特别在计算的开始阶段，当 $|a_{ij}|$ 很小时，变换收益甚少，浪费了计算时间，增加了舍入误差，所以，是不可取的。为了克服这一缺点，提出了所谓"过关雅可比法"。即在顺次地取出非零的 a_{ij} 时，凡是 $|a_{ij}|$ 小于某个"关值" α 时，就不去进行变换而继续找下一个 a_{ij}。同时，随着迭代过程的进展，α 也按一定规律减少，直到求得结果为止。这一方法的主要缺点在于"关值" α 不易选取。不过，由于一般情况下经过几次循环(我们称非对角线元素依次取完一遍为一个循环)，非对角线元素的数量级已比较接近。我们只需在前几个循环内采取过关的办法，以后即按循环方式处理。这样，"关值" α 的选取就较简单。最后这种方法是最常用的。

(ii) 减少舍入误差的影响

旋转法计算过程中的关键步骤是计算 $\sin\varphi$ 与 $\cos\varphi$ 的值。这对于计算结果的精确度有很大的影响。因而，必须尽可能准确地算出它们。考虑到矩阵元素各种可能的变化范围以及有效位消失等因素，采用如下公式较好：

$$
\begin{cases}
t=\dfrac{2a_{ij}}{a_{ii}-a_{jj}};\quad z=\dfrac{a_{ii}-a_{jj}}{2a_{ij}}\\
\cos 2\varphi=\begin{cases}(1+t^2)^{-1/2} & 若\ |t|<1\\ |z|\cdot(1+z^2)^{-1/2} & 若\ |t|\geqslant 1\end{cases}\\
\sin 2\varphi=\begin{cases}t\cdot(1+t^2)^{-1/2} & 若\ |t|<1\\ \operatorname{sign} z\cdot(1+z^2)^{-1/2} & 若\ |t|\geqslant 1\end{cases}\\
\cos\varphi=\left[\dfrac{1}{2}(1+\cos 2\varphi)\right]^{1/2}\\
\sin\varphi=\sin 2\varphi/2\cos\varphi
\end{cases}
\tag{10.3.36}
$$

另外，在进行变换时也常常采用一些措施来提高精度。例如，将计算对角线元素的公式换为如下等价形式：

$$
\begin{aligned}
a_{ii}^{(1)}&=a_{ii}+\operatorname{tg}\varphi\cdot a_{ij}\\
a_{jj}^{(1)}&=a_{jj}-\operatorname{tg}\varphi\cdot a_{ij}
\end{aligned}
\tag{10.3.37}
$$

每次变换时，相应对角线元的修正量 $\operatorname{tg}\varphi\cdot a_{ij}$ 分别用 n 个单元累加起来，完成一定次数变换(通常是非对角线元素跑完一遍，即一个循环)之后，再将其加到对角线元素上去。这样可以减少逐次变换的误差积累，有时，对提高小特征值的精度也有一定好处。非对角线元及特征向量的计算公式也可相应地换为如下等价形式：

$$
\left.
\begin{aligned}
a_{il}^{(1)}&=a_{il}+\sin\varphi\left(a_{jl}-\frac{\sin\varphi}{1+\cos\varphi}\,a_{il}\right)\\
a_{jl}^{(1)}&=a_{jl}-\sin\varphi\left(a_{il}+\frac{\sin\varphi}{1+\cos\varphi}\,a_{jl}\right)
\end{aligned}
\right\}\quad (l\neq i,\ j)
\tag{10.3.38}
$$

采取上述这些措施后，计算结果的精度将有所提高，计算时间也有所节省。许多计算表明，对于一般的实对称矩阵，求得最终结果所需总的旋转次数大约为 $3n^2\sim 5n^2$ 的数量级。结果精度差不多可达到所用的工作精度，只是很接近的特征值或绝对值较小的特征值有些误差。特征向量的精度可能不如特征值好，但其正交性总是较好的。

根据上述措施编制的一个旋转法计算程序，见本章最后所附的程序一。

(二)旋转法的推广

旋转法对于对称矩阵特征值问题的有效性是显著的，方法简单紧凑，精度高，收敛也较快。因此，自然联想到能否把它推广至其它类型矩阵的特征值问题。目前，这方面已经取得了一些结果，例如，对于正规矩阵(即满足条件：$\boldsymbol{AA}^T=\boldsymbol{A}^T\boldsymbol{A}$ 的矩阵)，旋转法稍加变形就可以直接应用，其效果也较好。但是，对于一般矩阵，直接应用旋转法求解却遇到了一定困难，长期没有取得重大进展。后来，人们从如下事实中得到启发，即对一般矩阵 $\boldsymbol{A}$ 而言，与其相似的所有矩阵 $\boldsymbol{P}^{-1}\boldsymbol{AP}$ 中，使 $N^2(\boldsymbol{P}^{-1}\boldsymbol{AP})$ 最小者，必为正规的($N^2(\boldsymbol{A})$ 表示矩阵 $\boldsymbol{A}$ 所有元素的平方和)。从这一事实出发，人们想到是否可能先用一系列相似变换来减少元素平方和并将矩阵 $\boldsymbol{A}$ 化为近似正规矩阵，然后用旋转法求这一近似正规矩阵的特征值和特征向量。这个想法目前已从不同途径得到实现，它们的基本思想都是把矩阵正规化和正规矩阵对角化这两个过程结合起来进行，所用的逐次变换矩阵则是旋转矩阵与某种初等矩阵的乘积。实践证明这些方法有一定效果，特别是对于单构矩阵(初等因子为线性的矩阵)，它们是很有效的。下面简单介绍一下这些方法。我们主要以实矩阵为例讨论其中的一种方法(见[11])来说明这类方法的基本特点。其他方法只简单地列举一下计算公式。

在§10.3.2中,我们已经讨论过初等变换矩阵$S_{ij}(\alpha)$的一些性质。从(10.3.6)式可以看出,适当选择参数α,可使变换后矩阵元素平方和减小(这是显然的,因为α充分小时,第一项占主导地位,只需令α与C_{ji}相同,(10.3.6)右端总是正数)。即是说,此时初等相似变换(10.3.4)将减少矩阵元素平方和。同时,我们又知道相似变换下矩阵元素平方和减至最小时,就得出正规矩阵。所以,我们可以期望,用变换(10.3.4)不断对矩阵$\boldsymbol{A}$进行变换,减小其元素平方和,直到其元素平方和减至最小,就得出与矩阵$\boldsymbol{A}$相似的正规矩阵,这样就可以完成矩阵$\boldsymbol{A}$的正规化过程。然后,再用旋转变换将这一正规矩阵对角化,就可以得出矩阵$\boldsymbol{A}$的特征值和特征向量。通常,为了计算方便起见,可将上述两个过程结合进行,即作一次初等变换(10.3.4),然后作一次旋转变换,如此交替进行,直至矩阵变为对角型为止。综上所述,我们可得如下计算格式:

$$\begin{cases} ①\ \tilde{\boldsymbol{A}}=S_{ij}(\alpha)\cdot\boldsymbol{A}\cdot S_{ij}(\alpha)^{-1} \\ ②\ \boldsymbol{A}_1=\boldsymbol{J}_{ij}(\varphi)^T\cdot\tilde{\boldsymbol{A}}\cdot\boldsymbol{J}_{ij}(\varphi) \\ ③\ \text{以 } \boldsymbol{A}_1 \text{ 代替 } \boldsymbol{A}\text{, 重复上式,直至 } \boldsymbol{A}_1 \text{ 收敛为止。} \end{cases} \tag{10.3.39}$$

逐次变换矩阵中的参数(i, j), α, φ可按下述办法选取:

足标$[i, j]$ $(i<j)$按行的顺序取值,即依次取$[1, 2]$, $[1, 3]$, …, $[1, n]$, $[2, 3]$, …, $[2, n]$, …, $[n-1, n]$。并如此循环。

α取为:

$$\alpha=C_{ji}/\max(|C_{ji}|,\ m_{ji}) \tag{10.3.40}$$

其中

$$m_{ji}=\sum_{k\neq j}a_{kj}^2+\sum_{k\neq i}a_{ik}^2+(a_{jj}-a_{ii})^2+9(a_{ij}^2+a_{ji}^2);$$

C_{ji}见(10.3.6)式。

φ值由下式决定:

$$\operatorname{tg} 2\varphi=\frac{\tilde{a}_{ij}+\tilde{a}_{ji}}{\tilde{a}_{ii}-\tilde{a}_{jj}} \qquad |\varphi|\leqslant\frac{\pi}{4} \tag{10.3.41}$$

已经证明,上述过程中$\boldsymbol{A}_1$将收敛于块对角型矩阵。特别是当矩阵$\boldsymbol{A}$是单构的且仅有实特征值时,$\boldsymbol{A}_1$将收敛于对角型矩阵,其渐近收敛速度是平方收敛的(即(i, j)取值一个循环后,非对角线元将由ε量级变为ε^2量级)。

实际计算中,逐次的α值会愈来愈小。若矩阵有重特征值则可能出现α值较大的情况,此时,相应的变换应该略去,否则会影响特征向量的精度。所以,通常按"过关"的办法来决定变换的取舍,α大于"关值",相应变换就略去。

上面,我们简单地介绍了一种广义的旋转法。类似方法还有几种,其基本思想均是相同的,仅是计算公式有所不同。例如,对于一般复矩阵适用的,使用所谓"剪切"变换的方法(见[3]),其计算公式为:

$$\begin{cases} \boldsymbol{A}^{(0)}=\boldsymbol{A} \\ \tilde{\boldsymbol{A}}^{(k)}=\boldsymbol{S}_{(k)}^{-1}\cdot\boldsymbol{A}^{(k)}\cdot\boldsymbol{S}^{(k)} \\ \boldsymbol{A}^{(k+1)}=\boldsymbol{U}^{(k)*}\cdot\tilde{\boldsymbol{A}}^{(k)}\cdot\boldsymbol{U}^{(k)} \end{cases} \quad (k=0,\ 1,\ 2,\ \cdots)$$

其中$\boldsymbol{S}^{(k)}$, $\boldsymbol{U}^{(k)}$除下列四个元素外,与单位矩阵相同。

$$\begin{pmatrix} S_{ii}^{(k)} & S_{ij}^{(k)} \\ S_{ji}^{(k)} & S_{jj}^{(k)} \end{pmatrix}=\begin{pmatrix} \cosh y_k, & -ie^{i\alpha k}\sinh y_k \\ ie^{-i\alpha k}\sinh y_k, & \cosh y_k \end{pmatrix}$$

$$\begin{pmatrix} U_{ii}^{(k)} & U_{ij}^{(k)} \\ U_{ji}^{(k)} & U_{jj}^{(k)} \end{pmatrix} = \begin{pmatrix} \cos\theta_k, & -e^{i\varphi_k}\sin\theta_k \\ e^{-i\varphi_k}\sin\theta_k, & \cos\theta_k \end{pmatrix}$$

又如，还有使用另一种相似变换代替“剪切”变换的方法见[14]，其计算格式与前者相同，只是其中 $S^{(k)}$ 为如下矩阵：

$$\begin{pmatrix} S_{ii}^{(k)} & S_{ij}^{(k)} \\ S_{ji}^{(k)} & S_{jj}^{(k)} \end{pmatrix} = \begin{pmatrix} 1-\sin\varphi_k\cdot t_k, & e^{i\alpha_k}(\cos\varphi_k+1)t_k \\ e^{-i\alpha_k}(\cos\varphi_k-1)t_k, & 1+\sin\varphi_k\cdot t_k \end{pmatrix}$$

上述这些方法的效果是相差不大的，它们对于单构矩阵都比较有效。但是对于具有非线性初等因子的矩阵，效果还不够理想。因而，它们还不能很有效地解决一般矩阵的特征值问题。目前，这方面的情况还在发展中。

为了使用方便，我们把[10]中采用“剪切”变换的计算程序附于本章后面(见程序 2)，有兴趣于使用的读者可以直接按程序说明使用，不必去细读程序。

10.3.5 化对称矩阵为三对角线型的方法

这个方法的基本思想是首先用正交相似变换把给定对称矩阵 $\boldsymbol{A}$ 化为对称的三对角线型矩阵，然后求解三对角线型矩阵的特征值问题。后一问题的求解较为简单，正交相似变换的三对角化过程也有较高的数值稳定性，所以，这一方法是有很多优点的。同时，它在仅需求特征值或少量特征向量时比旋转法更节省工作量，精确度也较高，是求解对称矩阵特征值问题的一个好方法。其主要缺点是算法和程序比较复杂，不如旋转法那样简单、紧凑。此外，由于必须先求出特征值，再用反幂法求相应特征向量。因而，需求全部或大部特征向量时，不如旋转法有利。

(一)化为三对角型

我们先讨论将对称阵 $\boldsymbol{A}$ 化为三对角型的问题。从 10.3.2 节最后的讨论知道，适当选择矩阵 $\boldsymbol{H}_1$，可以得到：

$$\boldsymbol{H}_1\cdot\boldsymbol{A}\cdot\boldsymbol{H}_1 = \begin{bmatrix} C_1, & x\cdots\cdots x \\ b_2, & x\cdots\cdots x \\ 0, & x\cdots\cdots x \\ \vdots & \vdots \qquad \vdots \\ 0, & x\cdots\cdots x \end{bmatrix}$$

其中

$$\boldsymbol{H}_1 = \boldsymbol{H}_1^T = \boldsymbol{I} - 2\boldsymbol{u}_1\boldsymbol{u}_1^T/(\boldsymbol{u}_1^T\cdot\boldsymbol{u}_1)$$

$$\boldsymbol{u}_1^T = (0,\ a_{21}-\alpha_1,\ a_{31},\ \cdots,\ a_{n1})$$

$$C_1 = a_{11}$$

$$b_1 = \pm\left(\sum_2^n a_{i1}^2\right)^{1/2}$$

由于矩阵 $\boldsymbol{A}$ 是对称矩阵，$\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1$ 也应是对称的。所以，其第一行元素也只有前两个元素不为零。这样一来矩阵 $\boldsymbol{A}_2 = \boldsymbol{H}_1\cdot\boldsymbol{A}\cdot\boldsymbol{H}_1$ 将有下列形式：

$$\boldsymbol{A}_2 = \begin{bmatrix} C_1, & b_2, & 0\cdots\cdots\cdots 0 \\ b_2, & a_{22}^{(2)}, & a_{23}^{(2)}\cdots\cdots a_{2n}^{(2)} \\ 0, & a_{32}^{(2)} & \cdots\cdots\cdots\cdots \quad \vdots \\ \vdots & \vdots & \\ 0, & a_{n2}^{(2)} & \cdots\cdots\cdots\cdots a_{nn}^{(2)} \end{bmatrix}$$

我们进一步考虑除去第一行和第一列的 $n-1$ 阶对称矩阵，按完全相同的办法可以用变换矩阵 $\widetilde{\boldsymbol{H}}_2$ 将其第一行和第一列化为类似形状。这相当于对 $\boldsymbol{A}_2$ 进行变换时，令变换矩阵 $\boldsymbol{H}_2$ 的第一行和第一列与单位矩阵相同，其它元素与 $\widetilde{\boldsymbol{H}}_2$ 相同而得的结果。因而，我们有：

$$\boldsymbol{A}_3=\boldsymbol{H}_2\boldsymbol{A}_2\boldsymbol{H}_2=\begin{bmatrix} C_1 & b_2 & 0 & 0\cdots\cdots 0 \\ b_2 & C_2 & b_3 & 0\cdots\cdots 0 \\ 0 & b_3 & a_{33}^{(3)}\cdots\cdots\cdots\cdots & a_{3n}^{(3)} \\ 0 & 0 & \vdots\cdots\cdots\cdots\cdots & \vdots \\ \vdots & \vdots & & \\ 0 & 0 & a_{n3}^{(3)}\cdots\cdots\cdots\cdots & a_{nn}^{(3)} \end{bmatrix}$$

其中

$$\boldsymbol{H}_2=\boldsymbol{I}-2\boldsymbol{u}_2\boldsymbol{u}_2^T/(\boldsymbol{u}_2^T\cdot\boldsymbol{u}_2)$$

$$\boldsymbol{u}_2^T=(0,\ 0,\ a_{32}^{(2)}-\alpha_2,\ a_{33}^{(2)},\ \cdots,\ a_{3n}^{(2)})$$

$$C_2=a_{22}^{(2)}$$

$$b_3=\pm\left(\sum_3^n a_{i2}^{(2)}\right)^{1/2}$$

如此继续下去，$n-2$ 步变换以后我们将得出一个三对角型矩阵：

$$\boldsymbol{A}_{n-1}=\boldsymbol{H}_{n-2}\cdot\boldsymbol{H}_{n-3}\cdot\cdots\cdot\boldsymbol{H}_1\cdot\boldsymbol{A}\cdot\boldsymbol{H}_1\cdot\boldsymbol{H}_2\cdot\cdots\cdot\boldsymbol{H}_{n-2}$$

$$=\begin{bmatrix} C_1 & b_2 & & & \\ b_2 & C_2 & b_3 & & \boldsymbol{0} \\ & b_3 & C_3 & b_4 & \\ & & \ddots & \ddots & b_n \\ \boldsymbol{0} & & & b_n & C_n \end{bmatrix}$$

这样就完成了矩阵 $\boldsymbol{A}$ 的三对角化过程。不难看出所有的矩阵 $\boldsymbol{A}_i$ 和 $\boldsymbol{H}_i$ 都是对称矩阵。

综上所述，我们可以得出如下计算公式：

$$\begin{cases} \boldsymbol{A}_1=\boldsymbol{A} \\ \boldsymbol{A}_{i+1}=\boldsymbol{H}_i\cdot\boldsymbol{A}_i\cdot\boldsymbol{H}_i \\ \boldsymbol{H}_i=\boldsymbol{I}-2\boldsymbol{u}_i\cdot\boldsymbol{u}_i^T/(\boldsymbol{u}_i^T\cdot\boldsymbol{u}_i) \\ \boldsymbol{u}_i^T=(\underbrace{0,0,\cdots,0}_{i\text{个零}},\ a_{i,i+1}^{(i)}-\alpha_i,\ a_{i,i+2}^{(i)},\ \cdots,\ a_{in}^{(i)}) \\ \alpha_i=-\operatorname{sign}(a_{i,i+1}^{(i)})\cdot\sqrt{a_{i,i+1}^{(i)2}+a_{i,i+2}^{(i)2}+\cdots+a_{in}^{(i)2}} \end{cases}\quad (i=1,\ 2,\ \cdots,\ n-2)\qquad(10.3.42)$$

为了节省计算工作量和编制程序方便，我们可以把 $\boldsymbol{A}_{i+1}$ 的表达式展开，并按计算的实际顺序将计算公式改写为：

(1) $S_i=(a_{i,i+1}^{(i)2}+a_{i,i+2}^{(i)2}+\cdots+a_{i,n}^{(i)2})^{1/2}$

(2) $\alpha_i=-\operatorname{sign}(a_{i,i+1}^{(i)})\cdot S_i$

(3) $h_i=\alpha_i^2-\alpha_i a_{i,i+1}^{(i)}$ $\left(即\ \dfrac{1}{2}\boldsymbol{u}_i^T\boldsymbol{u}_i\right)$

(4) $\boldsymbol{p}_i=\boldsymbol{A}_i\boldsymbol{u}_i/h_i$ $\qquad (i=1,\ 2,\ \cdots,\ n-2)\qquad(10.3.43)$

(5) $\boldsymbol{k}_i=\boldsymbol{u}_i^T\boldsymbol{p}_i/2h_i$

(6) $\boldsymbol{q}_i=\boldsymbol{p}_i-k_i\boldsymbol{u}_i$

(7) $\boldsymbol{A}_{i+1}=\boldsymbol{A}_i-\boldsymbol{u}_i\boldsymbol{q}_i^T-\boldsymbol{q}_i\boldsymbol{u}_i^T$

实际计算时，我们可将对称矩阵 $\boldsymbol{A}$ 上三角部分的元素按行排列地顺序存放，以作为原始数

据(即后面程序中的场 $A[1:n(n+1)/2]$)。逐次的变换矩阵 $\boldsymbol{H}_i$ 需要保留，因为求得三对角矩阵的某个特征向量 $\boldsymbol{z}$ 后，矩阵 $\boldsymbol{A}$ 的相应特征向量 $\boldsymbol{x}$ 由下式计算:

$$\boldsymbol{x}=\boldsymbol{H}_1\cdot\boldsymbol{H}_2\cdot\cdots\cdot\boldsymbol{H}_{n-3}\cdot\boldsymbol{H}_{n-2}\cdot\boldsymbol{z}$$

为此，我们只需将逐次的向量 $\boldsymbol{u}_i$ 保留在矩阵 $\boldsymbol{A}_{i+1}$ 中第 i 列的零元素位置上，这样 $\boldsymbol{H}_i$ 的存放并不占用额外的存储单元。求得的三对角型矩阵为:

$$\boldsymbol{A}_{n-1}=\begin{pmatrix} C_1, & b_2 & & & \\ b_2, & C_2, & b_3 & & \boldsymbol{0} \\ & b_3, & C_3, & b_4 & \\ & & \ddots & \ddots & b_n \\ \boldsymbol{0} & & & b_n & C_n \end{pmatrix} \tag{10.3.44}$$

我们将对角线元素 C_i 保存在场 $C[1:n]$ 中，将次对角线元 b_i 保留在场 $B[1:n]$ 中($B[1]$ 为任意数)，这些数据可直接作为求解三对角型特征值问题的初始数据。变换过程的某一步可能出现值 S_i^2 小于某个小量 EPS 的情况，此时，矩阵 $\boldsymbol{H}_i$ 的正交性将较差。我们可将变换 $\boldsymbol{H}_i\boldsymbol{A}_i\boldsymbol{H}_i$ 略去，这样对特征值的精度影响不大。EPS 的大小需视所用计算机而定。一般来说，如果机器字长为 t，其阶码允许变化范围为 $-q\sim+q$，且 q 比 t 大得多，则可令:

$$EPS=2^{-(q-t)}$$

对求得的 $\boldsymbol{A}_{n-1}$ 的特征向量进行变换时，我们假设其 n_1 至 n_2 个特征向量存于场 $Z[1:n, n_1:n_2]$ 中(Z 之每一列为一个特征向量)，回代程序对其进行变换，将求得的 $\boldsymbol{A}$ 的相应特征向量仍然存于场 Z 之内。

按上述方案编制的算法语言程序及程序的形式参数表见本章最后所附的程序三。

(二)对称三对角型矩阵的特征值问题

我们现在来讨论如何求得对称三对角型矩阵 $\boldsymbol{A}_{n-1}$ 的特征值和特征向量的问题。这一节中，施斗姆序列的性质将起重要作用，我们首先讨论它。然后再讨论如何计算特征值问题。

定义：闭区间 $[a, b]$ 上定义的实函数序列: $f_0(x)$, $f_1(x)$, $\cdots$, $f_{n-1}(x)$, $f_n(x)$，若满足如下条件，就称之为一个施斗姆(Sturm)序列:

(1) 函数 $f_i(x)$ $(i=0, 1, 2, \cdots, n)$ 均是连续函数。

(2) 函数 $f_0(x)$ 在区间 $[a, b]$ 上不变号。

(3) 序列中任何两个相邻的函数在区间 $[a, b]$ 内没有公共零点。

(4) 若函数 $f_i(x)$ $(i=1, 2, \cdots, n-1)$ 在区间 $[a, b]$ 内某点处为零，则相邻的两个函数 $f_{i-1}(x)$, $f_{i+1}(x)$ 在该点处应有相反符号。

(5) 在函数 $f_n(x)$ 的任一零点 $\bar{x}$ 处(属于区间 $[a, b]$ 的)，$f_{n-1}(x)\cdot f_n(x)$ 为降函数，即是

$$\operatorname{sign}[f_{n-1}(\bar{x}-h)\cdot f_n(\bar{x}-h)]=+1$$

和

$$\operatorname{sign}[f_{n-1}(\bar{x}+h)\cdot f_n(\bar{x}+h)]=-1$$

其中 h 为充分小的正数。

施斗姆序列有如下重要性质:

定理 3.3: 若 $f_0(x)$, $f_1(x)$, $\cdots$, $f_{n-1}(x)$, $f_n(x)$ 为一个施斗姆序列，$V(\xi)$ 为区间 $[a, b]$ 上某点 ξ 处，函数值序列 $f_0(\xi)$, $f_1(\xi)$, $\cdots$, $f_{n-1}(\xi)$, $f_n(\xi)$ 的同号数(或称为该函数序列在 $x=\xi$ 处的同号数)，则区间 $[a, b]$ 内函数 $f_n(x)$ 的零点数等于 $V(a)-V(b)$。

证明：我们考虑当ξ由a连续地变化至b时，函数值序列$f_0(\xi)$，$f_1(\xi)$，…，$f_{n-1}(\xi)$，$f_n(\xi)$的同号数变化的情况。由于$f_i(x)$为连续函数，所以，同号数仅当ξ通过$f_i(x)$的零点时才可能变化。由于$f_0(x)$不变号，故可分为下列两种情况来讨论：

(i) ξ通过序列内部某个函数$f_i(x)$ $(i=1, 2, \cdots, n-1)$的零点$\bar{x}$的情况。

此时，由施斗姆序列的性质(1)—(4)得知，h充分小时，只可能出现表10.1中的两种情况。

表 10.1

ξ	$f_{i-1}(\xi)$	$f_i(\xi)$	$f_{i+1}(\xi)$	ξ	$f_{i-1}(\xi)$	$f_i(\xi)$	$f_{i+1}(\xi)$
$\bar{x}-h$	+	±	−	$\bar{x}-h$	−	±	+
$\bar{x}$	+	0	−	$\bar{x}$	−	0	+
$\bar{x}+h$	+	∓	−	$\bar{x}+h$	−	∓	+

显然，无论那种情况下序列的同号数均不变。此外，若$\bar{x}$为$f_i(x)$的重零点，上述结论显然也是正确的。

(ii) ξ通过函数$f_n(x)$的零点$\bar{x}$的情况

此时，由施斗姆序列的性质(1)、(3)、(5)得知，h充分小时，只可能出现表10.2中的两种情况。

表 10.2

ξ	$f_{n-1}(\xi)$	$f_n(\xi)$	ξ	$f_{n-1}(\xi)$	$f_n(\xi)$
$\bar{x}-h$	−	−	$\bar{x}-h$	+	+
$\bar{x}$	−	0	$\bar{x}$	+	0
$\bar{x}+h$	−	+	$\bar{x}+h$	+	−

显然，无论那种情况，当ξ从$\bar{x}-h$增加至$\bar{x}+h$时序列均损失一个同号数。此外，若$\bar{x}$为$f_n(x)$的奇重零点，上述结论显然也是正确的。但应注意$\bar{x}$不可能为$f_n(x)$的偶重零点，因为这与施斗姆序列的性质(1)，(3)，(5)是矛盾的。

综上所述，当ξ由a连续地增加至b时㊀，序列$f_0(\xi)$，$f_1(\xi)$，…，$f_{n-1}(\xi)$，$f_n(\xi)$的同号数减少的数目，应等于区间$[a, b]$内部函数$f_n(x)$的零点数(奇重零点按单零点计算)。定理证完。

施斗姆序列的上述性质，在求解对称三对角型矩阵的特征值问题中有重要作用。因为对于对称三对角型矩阵，其逐次左上角主子矩阵的特征多项式将构成一个施斗姆序列。现在，我们就来证明这一事实。

考虑前面由正交相似变换所求得的对称三对角型矩阵$\boldsymbol{A}_{n-1}$。从计算的方便和有效考虑，我们可以假定矩阵$\boldsymbol{A}_{n-1}$是不可约的，即$b_i \neq 0$ $(i=2, 3, \cdots, n)$。因为若某个$b_i=0$，则可将$\boldsymbol{A}_{n-1}$分为两个三对角型矩阵的直接和，问题就化为两个低阶的特征值问题，并且，其中每一个都满足条件：$b_i \neq 0$。

㊀ 如果a恰为$f_n(\lambda)$的零点，则应取$f_n(a)$之符号与$f_{n-1}(a)$相反，以保证区间$[a, b]$内部的零点个数得以正确计数。同样，当b为$f_n(\lambda)$之零点，应使$f_n(b)$与$f_{n-1}(b)$同号。

矩阵 $\boldsymbol{A}_{n-1}$ 的特征多项式为:

$$\det(\boldsymbol{A}_{n-1}-\lambda \boldsymbol{I})=\begin{vmatrix} C_1-\lambda & b_2 & & & \boldsymbol{0} \\ b_2 & C_2-\lambda & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & & & b_n \\ \boldsymbol{0} & & & b_n & C_n-\lambda \end{vmatrix} \tag{10.3.45}$$

若以 $f_i(\lambda)$ 表示上面行列式左上角的 i 阶主子式, 用行列式展开的办法, 即可得出下列递推关系式:

$$\begin{cases} f_0(\lambda)=1 \\ f_1(\lambda)=C_1-\lambda \\ f_i(\lambda)=(C_i-\lambda)f_{i-1}(\lambda)-b_i^2 f_{i-2}(\lambda) \quad (i=2,\ 3,\ \cdots,\ n) \end{cases} \tag{10.3.46}$$

其中 $f_n(\lambda)$ 就是 $\boldsymbol{A}_{n-1}$ 的特征多项式。多项式序列:

$$f_0(\lambda),\ f_1(\lambda),\ \cdots,\ f_{n-1}(\lambda),\ f_n(\lambda)$$

显然满足施斗姆序列的条件(1), (2)。同时, 由于对称矩阵的性质, $f_{i-1}(\lambda)$ 的零点 $\lambda_k^{(i-1)}$ 必定与 $f_i(\lambda)$ 的零点 $\lambda_k^{(i)}$ 互相分隔地排列如下(参见[9]):

$$\lambda_i^{(i)}\leqslant\lambda_{i-1}^{(i-1)}\leqslant\lambda_{i-1}^{(i)}\leqslant\lambda_{i-2}^{(i-1)}\leqslant\cdots\leqslant\lambda_2^{(i)}\leqslant\lambda_1^{(i-1)}\leqslant\lambda_1^{(i)} \quad (i=2,\ 3,\ \cdots,\ n) \tag{10.3.47}$$

如果 $f_i(\lambda)$ 有重零点 $\bar{\lambda}$, 则 $\bar{\lambda}$ 亦必为 $f_{i-1}(\lambda)$ 的零点, 从递推关系式(10.3.46)得知 $\bar{\lambda}$ 必为 $f_{i-2}(\lambda)$ 的零点(因 $b_i\neq 0$)。如此递推下去, 将得出 $\bar{\lambda}$ 为 $f_0(\lambda)$ 的零点的结论, 显然这是不可能的。所以,序列中任一多项式均不可能有重零点。同样推理, 若 $f_i(\lambda)$ 与 $f_{i-1}(\lambda)$ 有公共零点,也发生矛盾。即是说序列中任何两个多项式均无公共零点, 从而, 性质(3)得以满足。亦即分隔关系(10.3.47)中只可能出现不等号。

从关系式(10.3.46)得知,若 $f_i(\bar{\lambda})=0$, 则有:

$$f_{i+1}(\bar{\lambda})=-b_{i+1}^2 f_{i-1}(\bar{\lambda})$$

因而, $f_{i+1}(\bar{\lambda})$ 与 $f_{i-1}(\bar{\lambda})$ 反号,这就是性质 (4)。

当 λ 充分大时, $f_n(\lambda)\sim(-1)^n\cdot\lambda^n$, $f_{n-1}(\lambda)\sim(-1)^{n-1}\lambda^{n-1}$。所以, $f_{n-1}(\lambda)\cdot f_n(\lambda)$ 小于零。当 λ 逐渐减小且通过 $f_n(\lambda)$ 的最大零点 $\lambda_1^{(n)}$ 时, 由于前面的分隔关系(10.3.47), 我们知道 $f_{n-1}(\lambda)$ 不变号而 $f_n(\lambda)$ 将改变符号, 即是说 $f_{n-1}(\lambda_1^{(n)}-h)\cdot f_n(\lambda_1^{(n)}-h)>0$。这样, $f_{n-1}(\lambda)\cdot f_n(\lambda)$ 在 $\lambda_1^{(n)}$ 处为降函数。此外, 通过 $\lambda_1^{(n-1)}$ 后 f_{n-1} 变号而 f_n 不变号, 故有 $f_{n-1}(\lambda_2^{(n)}+h)\cdot f_n(\lambda_2^{(1)}+h)$ 小于零。仿照前面的讨论,同样可以推知在 $\lambda_2^{(n)}$ 处 $f_{n-1}(\lambda)\cdot f_n(\lambda)$ 亦为降函数。这样, 利用连续性及分隔关系, 很容易由大至小地逐个说明在 $f_n(\lambda)$ 的零点 $\lambda_i^{(n)}$ 处, $f_{n-1}(\lambda)\cdot f_n(\lambda)$ 均为降函数。性质(5)也得以满足。所以,我们有如下定理:

定理 3.4: 对于不可约的对称三对角型矩阵(即所有 b_i 均不为零者), 其逐次左上角主子矩阵的特征多项式: $f_0(\lambda)=1$, $f_1(\lambda)$, $f_2(\lambda)$, $\cdots$, $f_n(\lambda)$ 构成一个施斗姆序列。

根据上述定理和前面已证明过的施斗姆序列的特性,很容易证明如下结果。

定理 3.5: 矩阵 $\boldsymbol{A}_{n-1}$ 的大于某个实数 μ 的特征值个数等于由递推关系式(10.3.46)所确定的多项式序列 $\{f_0(\lambda),\ f_1(\lambda),\ \cdots,\ f_{n-1}(\lambda),\ f_n(\lambda)\}$ 在 $\lambda=\mu$ 处的同号数 $V(\mu)$。

证明: 考虑区间 $[\mu,\ k]$, 其中 $k>\lambda_1^{(n)}$ 是一个充分大的正数。由于多项式序列中相邻两者在 $\lambda=k$ 总是反号, 故 $V(k)=0$。这样, 从定理 3.3 得知, 区间 $[\mu,\ k]$ 内部 $f_n(\lambda)$ 的零点数就应等于 $V(\mu)$。

利用这个定理，我们很容易设计一个计算矩阵 $\boldsymbol{A}_{n-1}$ 的特征值的方法。例如，我们来求 $\boldsymbol{A}_{n-1}$ 的由大到小顺序排列的第 m 个特征值 λ_m。

首先，我们找出一个区间 $[a, b]$，使其包含 λ_m（这一点很容易办到，因为 $\boldsymbol{A}_{n-1}$ 的全部特征值都包含在区间 $[-\max\limits_i(|b_{i-1}|+|c_i|+|b_i|),\ \max\limits_i(|b_{i-1}|+|c_i|+|b_i|)]$ 之内）。然后，计算区间中点 $\frac{a+b}{2}$ 处多项式序列 $\{f_0(\lambda),\ \cdots,\ f_n(\lambda)\}$ 的同号数 $V\left(\frac{a+b}{2}\right)$。如果 $V\left(\frac{a+b}{2}\right)<m$，则表明 λ_m 属于区间 $\left[a,\ \frac{a+b}{2}\right]$；否则，$\lambda_m$ 将属于区间 $\left[\frac{a+b}{2},\ b\right]$。这样，我们就可以把包含 λ_m 的区间缩短一半。重复这个过程，直到区间长度小于某个指定误差时，便得到 λ_m 的近似值（通常取最后一个区间的中点作为 λ_m）。这一过程就是所谓的区间分半法(bisection)，其收敛速度和数值稳定性均较高，对于求少量几个特征值是特别适宜的。此外，还可利用计算过程中的信息来缩短求后面特征值的初始区间之长度，使整个计算过程得到加速。因而，用这一方法来求大部或全部特征值，也是很有效的。

实际计算时，由于多项式 $f_i(\lambda)$ 的值变化较大，往往产生上溢或下溢的现象，使计算难以进行下去。为解决这一问题，最好是引入新的函数 $q_i(\lambda)=f_i(\lambda)/f_{i-1}(\lambda)\ (i=1,\ 2,\ \cdots,\ n)$。此时，递推公式(10.3.46)变为：

$$\begin{cases} q_1(\lambda)=c_1-\lambda \\ q_i(\lambda)=(c_i-\lambda)-b_i^2/q_{i-1}(\lambda) \quad (i=2,\ 3,\ \cdots,\ n) \end{cases} \tag{10.3.48}$$

显然，$V(\mu)$ 应等于 $q_i(\mu)\ (i=1,\ 2,\ \cdots,\ n)$ 中大于零的个数。因而，我们只需每次按递推公式(10.3.48)计算一下 $q_i(\mu)$ 的数值即可判断要求的特征值属于哪一个区间，整个计算过程便可进行下去。$q_i(\lambda)$ 的数值通常不太大，所以不会产生上溢的问题。但是，在计算过程中可能出现某个 $q_i(\lambda)$ 为零的情况，此时，我们可用一个正的小量 $\min s\sim 2^{-t}$（t 为计算机的字长）去代替它，使计算继续下去。这样作将保证 $q_{i+1}(\lambda)<0,\ q_{i+2}(\lambda)\sim c_{i+2}-\lambda$。这与实际情况是符合的，因而，不会影响求得结果的精度。

此外，如果总共要求计算 S 个相邻特征值，最好是由大至小顺序地求出，并用 $2S$ 个单元记录计算中间过程所得到的这些特征值上下界的信息，以缩短整个计算过程。

对于前节的三对角化方法所作的严格误差分析表明（见[9]及[6]卷 II），若采用标准浮点运算，三对角化所求得的矩阵 $\boldsymbol{A}_{n-1}$ 的准确特征值与原来矩阵 $\boldsymbol{A}$ 的特征值间的最大误差 ε_1 将满足如下不等式：

$$|\varepsilon_1|\leqslant\frac{f(n)\cdot 2^{-t}\|\boldsymbol{A}\|}{1-f(n)2^{-t}}$$

其中 $f(n)\sim 3.2n^{5/2}$。

然而，应该指出，某些实际计算的结果表明误差并没有那样大，一般说来最大误差 ε_1 仅为 $2^{-t+1}\|\boldsymbol{A}_{n-1}\|$。即是说绝对值较大的特征值只有最后少数几个二进制位上有误差。

对于区间分半法所作的严格误差分析（见[9]）表明，若采用标准浮点运算，则按区间分半法进行足够多次分半后，求得的特征值与 $\boldsymbol{A}_{n-1}$ 的准确特征值间的最大误差 ε_2 满足如下不等式：

$$|\varepsilon_2|\leqslant\rho\cdot 2^{-t}\|\boldsymbol{A}_{n-1}\|$$

其中 ρ 为常数，大小是 10 左右。

比较 ε_1 与 ε_2 的上界可以看出，一般来说，三对角化过程的误差是主要的，区间分半过程

的误差仅占次要地位。当 n 很大时，更是如此。

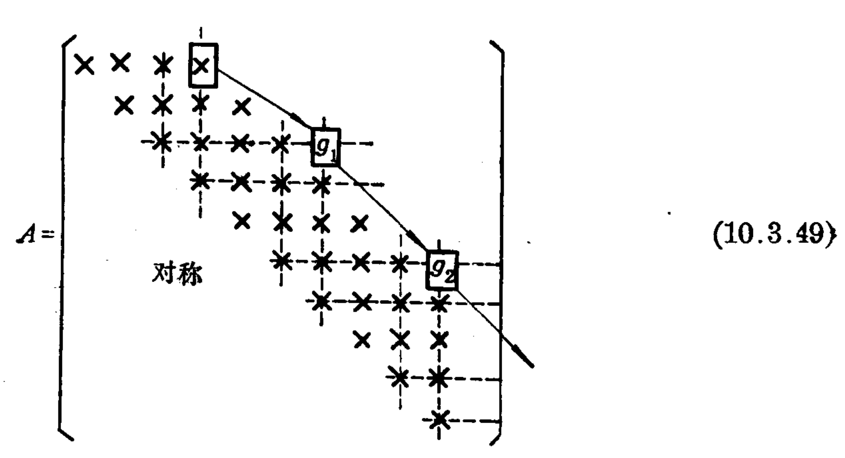
区间分半法的计算程序见本章最后所附的程序四。

求得 $\boldsymbol{A}_{n-1}$ 的特征值后，可以采用反幂法来求相应的特征向量。反幂法的计算过程已在10.3.3节中讨论过。这里不同的是矩阵是对称三对角型，计算过程相应地比较简单。因而，直接给出算法语言程序(见本章最后所附的程序五)。

(三)化对称带状矩阵为三对角型

实践中经常遇到对称带状矩阵的特征值问题。如果使用前两节的方法求解它们，由于不能利用带型这一特点，将浪费很多运算量及存储量，因而，有必要提出特殊的方法。这里采用的方法是先用一系列平面旋转矩阵构成的正交相似变换，在保持其带状特点的前提下，将矩阵化为三对角型。然后用前面的区间分半法和反幂法来求其特征值和特征向量。

现在来具体说明如何用一系列形如(10.3.10)(令 $\varphi=\psi$)的特殊旋转变换将带型对称矩阵 $\boldsymbol{A}$ 化为三对角型。为了理解方便，首先以如下10阶7对角线矩阵为例(由于对称，只列出其上三角部分)：



(10.3.49)

第一步，用旋转矩阵 $\boldsymbol{J}_{34}(\varphi_1)$ 按(10.3.10)式(令 $\varphi=\psi=\varphi_1$)进行变换，并选择 φ_1 使变换后的元素 $a_{14}^{(1)}$ 为零(由于 $a_{14}^{(1)}=-a_{13}\sin\varphi_1+a_{14}\cos\varphi_1$，故只需令 $\operatorname{ctg}\varphi_1=a_{13}/a_{14}$ 即可)。变换后第3行第7列处将出现一个非零元素 g_1(见(10.3.49))，其他元素仍保持在原来的"非零"带内。为了保持矩阵的带型特点，再用平面旋转矩阵 $\boldsymbol{J}_{67}(\varphi_2)$ 进行一次变换，并按同样办法选择 φ_2，使元素 g_1 化为零。这时，又将在第6行第10列处产生一个非零元素 g_2。再用平面旋转矩阵 $\boldsymbol{J}_{9,10}(\varphi_3)$ 进行变换，便可将此元素消去。这样就完成了把元素 a_{14} 化为零的变换过程。类似地，再用矩阵 $\boldsymbol{J}_{23}(\varphi_4)$，将元素 a_{13} 化为零并将出现在第2行第6列，第5行第9列处的元素顺次消去，就完成了化 a_{13} 为零的变换过程。此时，矩阵的第1行已成为要求的三对角形状。同样对第2行进行变换(此时第1行的形状将保持不变)，如此继续下去，最终得出要求的三对角型矩阵。这就是所采用的特殊旋转变换过程。

如果矩阵 $\boldsymbol{A}$ 是 $2m+1$ 对角型矩阵($2m+1<n$)，即其元素 a_{ij} 满足下列关系：

$$a_{ij}=0,\quad \text{当 } |i-j|>m$$

假定其前 $j-1$ 行已变换为三对角型，那么，变换第 j 行为要求形式的运算过程将如表10.3所示。

从表10.3可以看出，每次变换仅有相邻两行和两列发生变化，并且消除非零"带"内每一个元素的第一次变换与其后附加的变换在格式上是完全统一的。仅是决定旋转角的关系

表 10.3

j 行需消去的元素	第一次变换矩阵	第一个出现在带外的非零元 g 的位置	消除 g 所需的附加变换矩阵
$a_{j,j+k}$ 其中 $k=m, m-1, \cdots, 2$ 且 $j+k\leqslant n$	$J_{j+k-1,j+k}(\varphi)$	$(j+k-1, j+k+m)$ 其中 $j+k+m\leqslant n$	$J_{j+k+\mu m-1,j+k+\mu m}(\varphi)$ $\mu=1, 2, \cdots\left[\frac{n-k-j}{m}\right]$

式稍有差别。对于第一次变换，旋转角 φ 由下式确定：

$$a_{j,j+k}\cos\varphi-a_{j,j+k-1}\sin\varphi=0 \tag{10.3.50}$$

对于其后的附加变换，φ 角由下式决定：

$$g\cos\varphi-a_{j+k+(\mu-1)m-1,\ j+k+\mu m-1}\sin\varphi=0 \quad \left(\mu=1,\ 2,\ \cdots\left[\frac{n-k-j}{m}\right]\right)$$

考虑到这一差别后，很容易把各次变换在程序中统一起来。

上述过程相当于求得一个正交矩阵 $\boldsymbol{V}$，使得：

$$\boldsymbol{D}=\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}=\begin{bmatrix} C_1 & b_2 & & & \boldsymbol{0} \\ b_2 & C_2 & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & \boldsymbol{0} & & & b_n \\ & & & b_n & C_n \end{bmatrix}$$

其中矩阵 $\boldsymbol{V}$ 就是逐次变换矩阵 $\boldsymbol{J}_{p,q}(\varphi)$ 的乘积。我们可以用前面讨论过的区间分半法和反幂法求出矩阵 $\boldsymbol{D}$ 的特征值 λ_i 及相应特征向量 z_i。那么，$\boldsymbol{x}_i=\boldsymbol{V}\cdot\boldsymbol{z}_i$ 就是原来矩阵 $\boldsymbol{A}$ 相应于特征值 λ_i 的特征向量。

这里讨论的将带状对称矩阵化为三对角型的方法，用来求带状对称矩阵的特征值和特征向量是非常有效的。如果我们把每一行需要化为零的元素 a_{ij} 的个数近似地认为是 $(m-1)$，每一个元素所需的附加变换数目大约是 $(n-i)/m$，因而，化为三对角型所需完成的总的变换数 N 将满足如下不等式：

$$N\leqslant n^2\cdot(m-1)/2m$$

每次变换所需完成的乘法运算量大约是 $8m+13$ 次。因而，总的乘法量 N_m 应满足不等式：

$$N_m\leqslant n^2\cdot(m-1)\cdot(4+6.5/m)$$

当 $m\ll n$ 时，N_m 仅是 n^2 的数量级，故方法所需的运算量是很少的。

此外，误差分析的结果表明，通过计算所得的矩阵 $\boldsymbol{D}$ 的准确特征值 μ_i 与原来矩阵 $\boldsymbol{A}$ 的准确特征值 λ_i 满足如下关系（见[10] pp. 279）：

$$\left(\frac{\Sigma(\mu_i-\lambda_i)^2}{\Sigma\lambda_i^2}\right)^{1/2}\leqslant 12\cdot 2^{-t}\cdot n^{3/2}(1+6\times 2^{-t})^{4n-7}\cdot\frac{m-1}{m}$$

自然，实际计算的误差远比上式中的小。所以，可以认为 μ_i 与 λ_i 一般来说是非常接近的。总之，上面介绍的方法无论其在运算量和精确度方面都是较好的。对于带状对称矩阵，这是目前较好的方法之一。

这个方法的算法语言程序见本章最后所附的程序六。

10.3.6 广义代数特征值问题 $Ax=\lambda Bx$ 的解法

我们现在讨论无阻尼自由振动问题以及其它有关问题中所提出的广义代数特征值问题：$\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$。通常，矩阵 $\boldsymbol{A}$, $\boldsymbol{B}$ 均为对称矩阵，并且，$\boldsymbol{B}$ 是正定矩阵。我们分为两种情形来进行讨论，一种情况是 $\boldsymbol{A}$, $\boldsymbol{B}$ 均为阶数不很高的所谓"满矩阵"的情况，此时，矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的元素通常是逐个地存储在计算机的内存中的。另一种情况是 $\boldsymbol{A}$, $\boldsymbol{B}$ 均为阶数很高的所谓"稀疏矩阵"，其元素一般可以用简单的运算来产生，因而无须将它们存起来，在不能用简单运算产生时，通常也只能将矩阵的非零元素存放于内存。两种情况下所使用的方法是不同的。

(一)满矩阵的情形

由于此时矩阵 $\boldsymbol{A}$, $\boldsymbol{B}$ 的元素(无论它是否为零)是逐个地存在内存中的，所以，可使用对矩阵 $\boldsymbol{A}$, $\boldsymbol{B}$ 本身进行变换的方法。例如，由于 $\boldsymbol{B}$ 是正定矩阵，我们可用平方根法将其分解为下三角型矩阵 $\boldsymbol{L}$ 与其转置矩阵 $\boldsymbol{L}^T$ 的乘积(参见 §8.1)。

$$\boldsymbol{B}=\boldsymbol{L}\cdot\boldsymbol{L}^T \quad (\boldsymbol{L}\text{ 为非奇下三角型矩阵})$$

因而有：

$$\boldsymbol{Ax}=\lambda\boldsymbol{L}\cdot\boldsymbol{L}^T\boldsymbol{x}$$

或者：

$$(\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{L}^{-T})(\boldsymbol{L}^T\boldsymbol{x})=\lambda(\boldsymbol{L}^T\boldsymbol{x}),\quad \boldsymbol{L}^{-T}=(\boldsymbol{L}^{-1})^T \tag{10.3.51}$$

由上式可知，广义代数特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$ 的求解可以化为对称矩阵 $\boldsymbol{Y}=\boldsymbol{L}^{-1}\cdot\boldsymbol{A}\cdot\boldsymbol{L}^{-T}$ 的特征值问题。矩阵 $\boldsymbol{Y}$ 的特征值 λ_i 就是要求的特征值，矩阵 $\boldsymbol{Y}$ 的特征向量 $\boldsymbol{z}_i$ 左乘以矩阵 $\boldsymbol{L}^{-T}$ 就是要求的特征向量 $\boldsymbol{x}_i=\boldsymbol{L}^{-T}\boldsymbol{z}_i$。

实际计算时，我们通常将矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的上三角部分以按行排列顺序存放于场 $A, B[1:n(n+1)/2]$。计算的第一步是分解矩阵 $\boldsymbol{B}$，将求得的 $\boldsymbol{L}^T$ 按同样方式存于场 B。第二步是解方程式：$\boldsymbol{X}\cdot\boldsymbol{L}^T=\boldsymbol{A}$，求得的解答 $\boldsymbol{X}=\boldsymbol{A}\cdot\boldsymbol{L}^{-T}$ 的下三角部分(注意：以后的计算只需用到 $\boldsymbol{X}$ 的下三角部分)按列排列存放于场 A。第三步是解方程式：$\boldsymbol{LY}=\boldsymbol{X}$，求出的解答：$\boldsymbol{Y}=\boldsymbol{L}^{-1}\cdot\boldsymbol{A}\cdot\boldsymbol{L}^{-T}$ 的上三角部分按行排列存放于场 A。第四步就是引用前述的任一种解对称矩阵特征值问题的程序，求出矩阵 $\boldsymbol{Y}$ 的特征值 λ_i 及特征向量 $\boldsymbol{z}_i$。最后，解方程式 $\boldsymbol{L}^T\boldsymbol{x}_i=\boldsymbol{z}_i$ 即得出要求的结果 $\boldsymbol{x}_i$。

上述方法一般说来精度是很高的；与其它方法比较，工作量也较小。但是，应该指出，当矩阵 $\boldsymbol{B}$ 对求逆来说是"病态"矩阵时，此法精度差一些(这种情况在矩阵阶数不高时较少遇到)。

按照这个方法把广义特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$ 化为普通的特征值问题的算法语言程序见本章最后所附的程序七。

还有一种较为简单的办法也可用来解广义代数特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$。其过程归结为两次使用旋转法。

首先用旋转法求得矩阵 $\boldsymbol{B}$ 的全部特征值 μ_i 和特征向量 $\boldsymbol{V}_i$。于是我们有：

$$\boldsymbol{B}=\boldsymbol{VDV}^T$$

其中 $\boldsymbol{V}$ 为特征向量列 $\boldsymbol{V}_i$ 所构成的正交矩阵，$\boldsymbol{D}$ 为对角型矩阵：$\mathrm{diag}(\mu_1, \mu_2, \cdots, \mu_n)$。

然后，按上式代替 $\boldsymbol{B}$，我们有：

$$\boldsymbol{Ax}=\lambda\boldsymbol{VDV}^T\boldsymbol{x}$$

因为 $\mu_i>0$，所以：$\boldsymbol{D}=\sqrt{\boldsymbol{D}}\cdot\sqrt{\boldsymbol{D}}$。$\sqrt{\boldsymbol{D}}=\mathrm{diag}(\sqrt{\mu_1},\ \cdots,\ \sqrt{\mu_n})$。于是，上式可写为：

$$\sqrt{\boldsymbol{D}}^{-1}\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}\sqrt{\boldsymbol{D}}^{-1}(\sqrt{\boldsymbol{D}}\boldsymbol{V}^T\boldsymbol{x})=\lambda(\sqrt{\boldsymbol{D}}\boldsymbol{V}^T\boldsymbol{x})$$

再对矩阵 $\boldsymbol{Y}=\sqrt{\boldsymbol{D}}^{-1}\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}\sqrt{\boldsymbol{D}}^{-1}$ 使用一次旋转法，求出其全部特征值 λ_i 和特征向量 $\boldsymbol{z}_i$，那么，λ_i 和 $\boldsymbol{x}_i=\boldsymbol{V}\sqrt{\boldsymbol{D}}^{-1}\boldsymbol{z}_i$ 就是要求的结果。

这一方法的工作量较前面讨论过的方法大一些，所需存储量也较大，但程序简单，数值稳定性较好（精度较高），也是一种可取的方法。特别是当矩阵 $\boldsymbol{B}$ 对求逆是“病态”的情况（此时矩阵 $\boldsymbol{D}$ 的某些对角线元很小），特征值 λ_i 中的某些（例如，中等大小者），仍能较准确地求得。这是前一方法所不及之处。

最后，还应指出，若矩阵 $\boldsymbol{B}$ 是正半定的，前面两种方法将失效。这种情况也有专门的处理方法，不过，由于它在实践中出现的机会少一些，我们不再涉及。有兴趣的读者可以参阅[15, 16]。

（二）高阶稀疏矩阵的情形

这是实践中最常遇到的情况。例如，用有限元法或有限差分法求解弹性结构的振动问题时，矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的阶数可以高达几千阶，而其非零元素往往只有1%左右。这类问题的主要困难在于阶数太高，目前的计算机无论从速度与存储容量方面都不能满足使用适合于低阶情况的方法来直接求解它们的要求。因而必须采取特殊的方法。这里选出少量方法来作一些简单介绍，详细情况读者可以参阅[17, 18, 19]。

(1) 区间分半法

这一方法适用于矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 均为带型且 $\boldsymbol{B}$ 为正定矩阵的情况。其基本思想与前面讨论过的求对称三对角型矩阵特征值的区间分半法类似。

我们知道，若 $\boldsymbol{P}$ 为任意实对称矩阵，则其逐次左上角主子矩阵的特征值之间有形如(10.3.47)的分隔关系。如果这一分隔关系中只有不等号出现，那么，对称三对角型矩阵的结论（见10.3.5节的定理3.4，定理3.5），对于矩阵 $\boldsymbol{P}$ 仍然是成立的。即是说，此时矩阵 $\boldsymbol{P}$ 大于某个实数 μ 的特征值个数应等于序列：$p_0(\mu)=1$，$p_1(\mu)=p_{11}-\mu$，$\cdots$，$p_{n-1}(\mu)$，$p_n(\mu)=\det(\boldsymbol{P}-\mu\boldsymbol{I})$ 的同号数。其中 $p_i(\mu)$ 为 $\det(\boldsymbol{P}-\mu\boldsymbol{I})$ 的左上角 i 阶主子式。

此外，前面已经证明当 $\boldsymbol{B}$ 为正定矩阵时，$\det(\boldsymbol{A}-\lambda\boldsymbol{B})=0$ 的根与对称矩阵 $\boldsymbol{P}=\boldsymbol{L}^{-1}\cdot\boldsymbol{A}\cdot\boldsymbol{L}^{-T}$ 的特征值是相同的。所以，求出对称矩阵 $\boldsymbol{P}$ 的特征值便得到所要求的结果。这样，就完全可以按照三对角型矩阵的区间分半法来计算广义特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$ 的特征值 λ_i。自然，如果直接算出矩阵 $\boldsymbol{P}$，然后计算序列 $p_0(\mu)$，$p_1(\mu)$，$\cdots$，$p_n(\mu)$ 的同号数，进而确定特征值所在区间等等，是会遇到很大困难的。因为矩阵 $\boldsymbol{P}=\boldsymbol{L}^{-1}\cdot\boldsymbol{A}\cdot\boldsymbol{L}^{-1}$ 一般来说是满矩阵，计算机的内存将容纳不下。利用如下结果，可以顺利地解决这一问题。

定理3.6：如果 $\boldsymbol{B}$ 为正定矩阵，则 $\det(\boldsymbol{A}-\lambda\boldsymbol{B})$ 的逐次左上角主子式 $\det(\boldsymbol{A}_r-\lambda\boldsymbol{B}_r)$ 的符号，与 $\det(\boldsymbol{P}-\lambda\boldsymbol{I})$ 的逐次左上角主子式 $p_r(\lambda)$ 的符号相同。

证明：因为矩阵 $\boldsymbol{B}$ 是正定矩阵，故有如下分解式：

$$\boldsymbol{B}=\boldsymbol{L}\boldsymbol{L}^T$$

可将 $\boldsymbol{L}$ 按下列形式分块：

$$L=\left(\begin{array}{c|c} \overbrace{L_r}^{r} & 0 \\ \hline M_r & N_r \end{array}\right)\Big\}r$$

直接验证容易得知:

$$B_r=L_r\cdot L_r^T \quad (B_r\text{ 为 }B\text{ 的左上角 }r\text{ 阶主子块})$$

$$L^{-1}=\begin{pmatrix} L_r^{-1} & 0 \\ X & Y \end{pmatrix}$$

$$P_r=L_r^{-1}A_rL_r^{-T} \quad (P_r,\ A_r\text{ 分别为矩阵 }P\text{ 及 }A\text{ 的左上角 }r\text{ 阶主子块})$$

于是,我们有:

$$\det(A_r-\lambda B_r)=\det(L_r[L_r^{-1}A_rL_r^{-T}-\lambda I]L_r^T)=(\det L_r)^2\cdot\det(P_r-\lambda I)$$

由于第一项恒正,所以, $\det(A_r-\lambda B_r)$与$\det(P_r-\lambda I)$同号。

根据这一结果,我们只需求出序列$\{\det(A_r-\mu B_r)\}_{r=1}^n$ 的符号,即可判定对称矩阵P因而问题 $Ax=\lambda Bx$ 的大于 μ 的特征值个数,区间分半法就可以进行下去。由于 A, B 均为带型,求出 $\det(A_r-\mu B_r)$的符号比较容易,而且,仅用内存也可处理很高阶的情况。这就使得区间分半法能够有效地解决高阶问题。

一般采用局部列主元素消去法来求出序列 $\{\det(A_r-\mu B_r)\}_{r=1}^n$的符号。为叙述方便起见,我们假定矩阵 A, B 的带宽均为 $2m+1$, 并认为其元素可以用简单公式产生或已存放至外存可以随时按行取入内存,且内存可以存放 $m+1$ 行以上的非零元素,即其容量大于 $(m+1)\times(2m+1)$。此外,用(i, j)表示处于内存第 i 行第 j 列处的元素。

首先,把 $A-\mu B$ 的前 $m+1$ 行调入内存,并用局部列主元素消去法将其消为上三角型。具体来说,先考察$|(2, 1)|$即$|a_{21}-\mu b_{21}|$是否大于$|(1, 1)|$即$|a_{11}-\mu b_{11}|$,若大于,则交换一、二两行,然后将元素$(2, 1)$消去。这样,二阶主子式的符号即可得出(应考虑进行交换后行列式改变符号),第二行就消去了。对于第三行消去时,先比较$|a_{31}-\mu b_{31}|$(即$|(3, 1)|$)与$|a_{11}^*-\mu b_{11}^*|$的大小并作必要的行交换和消去元素$(3, 1)$;然后再比较元素$(2, 2)$与元素$(3, 2)$之大小决定是否要交换二、三两行,并消去$(3, 2)$元素。这时三阶主子式之符号也已求得,第三行也就消去了。如此继续对每一行进行消去,直至 $m+1$ 行消去完毕,$m+1$ 阶主子式的符号就算求得。

从 $m+2$ 行起,每行的消去按如下步骤进行:

(i) 将内存中的全部 $m+1$ 行元素向上移一行,向左移一列(即使得$(2, 2)$元素处于$(1, 1)$元素的位置)。

(ii) 调入 $m+2$ 行(或紧接于内存最末行之后的行)至内存 $m+1$ 行位置上。

(iii) 若元素$(m+1, i)$按模大于元素(i, i),则交换 $m+1$ 行与第 i 行。

(iv) 用第 i 行消去 $m+1$ 行的第 i 列元素,即消去元素$(m+1, i)$。

对于 $i=1, 2, \cdots, m$ 重复(iii)、(iv),即将调入的 $m+2$ 行(或紧接于最后的行)消去完毕。此时有:

$$\operatorname{sign} p_{m+2}(\mu)=(-1)^N\cdot S\cdot\operatorname{sign}(a_{m+1,m+1}^*)$$

其中 N 为消去 $m+2$ 行中的交换次数,S 为前一子式的符号,$a_{m+1,m+1}^*$为消去完毕后$(m+1, m+1)$位置上的元素。

显然，按上述格式进行消去，无论阶数 n 有多大，内存中一般只需 $(m+1)\cdot(m+2)$ 个工作单元。因而，其节省存储量的优点是十分显著的。此外，所需运算量也较少，大约仅需完成 $2m^2n\times NR\times F$ 次乘法，NR 为要求的根的个数，F 为每根的迭代次数。当 m 远小于 n 时，其运算量也较节省。这个方法的精确度一般也较高，并且不受矩阵 $\boldsymbol{B}$ 对求逆病态的影响。其使用上也较灵活，求各特征值时互不依赖等等。因而，是一个可行的方法。

然而，使用上述方法不能同时求得特征向量，必须在求得特征值后，再用反幂法(详见后面)求其相应特征向量。这是这个方法的缺点之一。另外，如果逐次左上角主子式中有一些为零，特别是相邻几个主子式同时为零时，计算会发生困难。不过，由于矩阵元素本身及计算过程中舍入误差的存在，通常很少遇到这种情况。

(2) 反幂法

我们知道，广义特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$ 可以写为等价形式：

$$\boldsymbol{Py}=\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{L}^{-T}(\boldsymbol{L}^T\boldsymbol{x})=\lambda(\boldsymbol{L}^T\boldsymbol{x})=\lambda\boldsymbol{y}$$

如果知道某个特征值的近似值 μ，则用反幂法求其相应特征向量的迭代公式可写为：

$$(\boldsymbol{P}-\mu\boldsymbol{I})\boldsymbol{y}^{(k+1)}=q_{k+1}\boldsymbol{y}^{(k)}\quad(k=0,\ 1,\ 2,\ \cdots)$$

其中 q_{k+1} 为某个比例因子，将其引入的目的在于使 $\|\boldsymbol{y}_{k+1}\|$ 不致过大。

将上式中的 $\boldsymbol{P}$ 及 $\boldsymbol{y}$ 还原为 $\boldsymbol{A}$, $\boldsymbol{B}$ 及 $\boldsymbol{x}$, 我们就得出如下反幂法的迭代公式：

$$(\boldsymbol{A}-\mu\boldsymbol{B})\cdot\boldsymbol{x}^{(k+1)}=q_{k+1}\boldsymbol{B}\boldsymbol{x}^{(k)}\quad(k=0,\ 1,\ 2,\ \cdots)$$

用反幂法求特征向量需要每步解一个代数方程组，由于矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都是高阶的稀疏矩阵，我们可以使用本书第一章中所讨论的一些方法求解，以节省运算量和提高能够处理的阶数。如果近似值 μ 的误差不大，一般迭代少数几次就能获得结果，这一点与前面讨论过的反幂法是相同的。

如果不知道特征值的近似值，我们也可以用逐次的迭代向量去估计一个近似值，而将迭代过程改为如下形式：

(i) 任选一个初始向量 $\boldsymbol{x}^{(0)}$;

(ii) 将 $\boldsymbol{x}^{(k)}$ 归一化，即使得 $\boldsymbol{x}^{(k)}$ 满足：

$$\boldsymbol{x}^{(k)T}\boldsymbol{B}\boldsymbol{x}^{(k)}=1;$$

(iii) 计算 $\mu_k=\boldsymbol{x}^{(k)T}\boldsymbol{A}\boldsymbol{x}^{(k)}$;

(iv) 解下列方程组求出 $\boldsymbol{x}^{(k+1)}$:

$$(\boldsymbol{A}-\mu_k\boldsymbol{B})\boldsymbol{x}^{(k+1)}=q_{k+1}\boldsymbol{B}\boldsymbol{x}^{(k)};$$

(v) 进行收敛条件检验。若未收敛，则以 $\boldsymbol{x}^{(k+1)}$ 代替 $\boldsymbol{x}^{(k)}$ 重复(ii), (iii), (iv)。

可以证明上述方程中 μ_k 收敛于某个特征值，$\boldsymbol{x}^{(k)}$ 收敛于相应特征向量。这一方法又称为瑞利(Rayleigh)商迭代法。其收敛速度较快，一般来说，它是立方收敛的。即如果 μ_k 与某个特征值 λ_i 之差为 $O(\varepsilon)$ 级小量，则 μ_{k+1} 与 λ_i 之差为 $O(\varepsilon^3)$ 级的小量。但是，这一方法的计算量还是较大的，并且比较难于控制 μ_k 收敛于某个需要求出的特征值，因为预先无法知道初始向量应该怎样选取。此外，若已求得一个特征向量，以后的初始向量及逐次迭代向量必须与其 $\boldsymbol{B}$-正交才能保证求得新的特征向量，这样也会增加许多运算量和存储量。因而，这一方法的实际使用效果将受到较大限制。

目前，反幂法使用于广义特征值问题上最成功的形式是所谓同时反(幂)迭代法、即对矩阵 $\boldsymbol{B}^{-1}\cdot\boldsymbol{A}$ 进行同时反迭代(参见 10.3.3 节第 3 段)，每步保持迭代向量的 $\boldsymbol{B}$-正交性，其具

体计算步骤如下:

(a) 任取 $n\times p$ 矩阵 $\boldsymbol{X}_0$(其列为 $\boldsymbol{B}$-正交归一向量)为初始矩阵。

(b) 解方程组:

$$\boldsymbol{A}\boldsymbol{Y}_{k+1}=\boldsymbol{B}\cdot\boldsymbol{X}_k$$

求出矩阵 $\boldsymbol{Y}_{k+1}$。

(c) 计算 $\boldsymbol{G}_{k+1}=\boldsymbol{Y}_{k+1}^T\cdot\boldsymbol{B}\cdot\boldsymbol{Y}_{k+1}$ 及 $\boldsymbol{H}_{k+1}=\boldsymbol{Y}_{k+1}^T\cdot\boldsymbol{A}\cdot\boldsymbol{Y}_{k+1}$

(d) 解 p 阶广义特征值问题:

$$\boldsymbol{H}_{k+1}\cdot\boldsymbol{Q}_{k+1}=\boldsymbol{G}_{k+1}\cdot\boldsymbol{Q}_{k+1}\cdot\boldsymbol{v}_{k+1}$$

求出 $p\times p$ 矩阵 $\boldsymbol{Q}_{k+1}$ 及对角线矩阵 $\boldsymbol{v}_{k+1}$。

(e) 计算 $k+1$ 次近似解 $\boldsymbol{X}_{k+1}$:

$$\boldsymbol{X}_{k+1}=\boldsymbol{Y}_{k+1}\cdot\boldsymbol{Q}_{k+1}$$

对于 $k=0,1,2,\cdots$, 重复(b)～(e), 直至 $\boldsymbol{X}_k$ 收敛为止。$\boldsymbol{X}_k$ 之各列即为要求的特征向量,$\boldsymbol{v}_k$ 之对角线元即为相应的特征值。一般来说,可用(10.3.5)式的方法或用两次雅可比法求解(d),若求 l 个最小的特征值及相应的特征向量,可令 $p=\min(2l, l+8)$。通常此法效果均较好。

(3) 极小化法

广义代数特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$ 的求解与求其相应瑞利(Rayleigh)商函数

$$f=(\boldsymbol{Ax}, \boldsymbol{x})/(\boldsymbol{Bx}, \boldsymbol{x})$$

的极小值和极小化向量问题是等价的(见[2]第十章§7)。因而, 可以使用求函数极小值的方法来计算要求的特征值和特征向量,基于这种思想的方法,统称为极小化方法。目前已经提出了一些这一类的计算方法, 这里仅举出其中一种来说明这些方法的特点, 详细的讨论, 读者可以参阅[19]。

我们知道, 对于求解对称正定矩阵的线性代数方程组 $\boldsymbol{Ax}=\boldsymbol{b}$, 共轭斜量法是一个有效的方法。在那里,要求极小化的二次函数是

$$f=\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Ax}-\boldsymbol{x}^T\boldsymbol{b}$$

其计算步骤为(参阅本书§8.2):

(i) 任选一个初始向量 $\boldsymbol{x}_0$,计算出函数 f 在 $\boldsymbol{x}_0$ 处的梯度 $\boldsymbol{r}_0$, 并令 $\boldsymbol{p}_0=\boldsymbol{r}_0$。

(ii) 按如下的公式从 i 次近似向量 $\boldsymbol{x}_i$ 计算出 $i+1$ 次近似向量 $\boldsymbol{x}_{i+1}$:

$$\begin{aligned}&\boldsymbol{x}_{i+1}=\boldsymbol{x}_i+\alpha_i\boldsymbol{p}_i\\&\boldsymbol{r}_{i+1}=-\operatorname{grad}f|_{x=x_{i+1}}=\boldsymbol{b}-\boldsymbol{Ax}_{i+1}\\&\boldsymbol{p}_{i+1}=\boldsymbol{r}_{i+1}+\beta_i\cdot\boldsymbol{p}_i\end{aligned}$$

其中 α_i, β_i 由分别使 f 在 $\boldsymbol{x}_{i+1}$ 达极小值以及所谓共轭关系来确定。

(iii) 对于 $i=1, 2, \cdots$。重复(ii),直至 $r_i\approx0$, 则 $\boldsymbol{x}_i$ 为所求的解向量。

上述共轭斜量法的计算格式已被推广至非二次函数的情况, 并取得很好效果。当 $f=\boldsymbol{x}^T\boldsymbol{Ax}/\boldsymbol{x}^T\boldsymbol{Bx}$ 时,自然亦可建立类似的计算格式,由于此时:

$$\operatorname{grad}f=\frac{2}{\boldsymbol{x}^T\boldsymbol{Bx}}(\boldsymbol{Ax}-\mu\boldsymbol{Bx})$$

其中 μ 为函数 f 在点 x 的值。故梯度方向与如下剩余向量一致:

$$\boldsymbol{r}=\boldsymbol{A}\boldsymbol{x}-\mu\boldsymbol{B}\boldsymbol{x}$$

如果我们采取简化办法计算 α_i, β_i(即不去求使 f 在 $\boldsymbol{x}_{i+1}$ 上达极小的准确 α_i, 及精确满足共轭关系式的 β_i 值,而用一种简化公式代替之),我们就可以把计算公式写为:

(i) 选取一个初始向量 $\boldsymbol{x}_0$, 并令:

$$\mu_0=\boldsymbol{x}_0^T\boldsymbol{A}\boldsymbol{x}_0/\boldsymbol{x}_0^T\boldsymbol{B}\boldsymbol{x}_0$$

$$\boldsymbol{r}_0=\boldsymbol{A}\boldsymbol{x}_0-\mu_0\boldsymbol{B}\boldsymbol{x}_0$$

$$\boldsymbol{p}_0=\boldsymbol{r}_0$$

(ii) 计算 $\alpha_i=\dfrac{-2}{a_i d_i+\sqrt{\Delta_i}}$

其中:

$$a_i=\frac{\boldsymbol{p}_i^T\boldsymbol{B}\boldsymbol{p}_i}{\boldsymbol{r}_i^T\boldsymbol{r}_i};\quad b_i=\frac{\boldsymbol{p}_i^T\boldsymbol{B}\boldsymbol{x}_i}{\boldsymbol{x}_i^T\boldsymbol{B}\boldsymbol{x}_i};\quad c_i=\frac{\boldsymbol{p}_i^T\boldsymbol{B}\boldsymbol{p}_i}{\boldsymbol{x}_i^T\boldsymbol{B}\boldsymbol{x}_i}$$

$$d_i=\frac{\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_i}{\boldsymbol{p}_i^T\boldsymbol{B}\boldsymbol{p}_i}-\frac{\boldsymbol{x}_i^T\boldsymbol{A}\boldsymbol{x}_i}{\boldsymbol{x}_i^T\boldsymbol{B}\boldsymbol{x}_i};\quad \Delta_i=a_i^2d_i^2-4(a_ib_id_i-c_i)$$

(iii) $\boldsymbol{x}_{i+1}=\boldsymbol{x}_i+\alpha_i\boldsymbol{p}_i$

(iv) $\mu_{i+1}=\boldsymbol{x}_{i+1}^T\boldsymbol{A}\boldsymbol{x}_{i+1}/\boldsymbol{x}_{i+1}^T\boldsymbol{B}\boldsymbol{x}_{i+1}$

(v) $\boldsymbol{r}_{i+1}=\boldsymbol{A}\boldsymbol{x}_{i+1}-\mu_{i+1}\boldsymbol{B}\boldsymbol{x}_{i+1}$

(vi) $\beta_i=\boldsymbol{r}_{i+1}^T\boldsymbol{r}_{i+1}/\boldsymbol{r}_i^T\boldsymbol{r}_i,\qquad \boldsymbol{p}_{i+1}=\boldsymbol{r}_{i+1}+\beta_i\boldsymbol{p}_i$

(vii) 对于 $i=0, 1, 2, \cdots$。重复(ii～vi),直至收敛为止。此时, μ 及 $\boldsymbol{x}$ 即要求的相应结果。

上述的极小化方法一定条件下将收敛至最小特征值 λ 及相应特征向量 $\boldsymbol{v}$。如果还需求出其它的特征值,则应将函数 f 改为下列形式,并依此类推:

$$f=\boldsymbol{x}^T\boldsymbol{p}^T\boldsymbol{A}\boldsymbol{p}\boldsymbol{x}/\boldsymbol{x}^T\boldsymbol{p}^T\boldsymbol{B}\boldsymbol{p}\boldsymbol{x}$$

其中, $\boldsymbol{p}=\boldsymbol{I}-\boldsymbol{v}\boldsymbol{v}^T\cdot\boldsymbol{B}/\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}$。或者,附加以正交性条件,变为约束极小值问题来处理。

第(vi)中计算 β_i 的公式还可以换为如下更复杂一些的形式:

$$\beta_i=-\boldsymbol{p}_i^T(\boldsymbol{A}-\mu_{i+1}\boldsymbol{B})\boldsymbol{r}_{i+1}/\boldsymbol{p}_i^T(\boldsymbol{A}-\mu_{i+1}\boldsymbol{B})\boldsymbol{p}_i$$

此时,收敛速度将有所提高。但因(vi)中的公式简单,也是可取的。

从上述讨论可知,极小化方法的每一步需要完成的主要工作为 $\boldsymbol{A}\boldsymbol{x}$ 或 $\boldsymbol{B}\boldsymbol{x}$ 类型的运算。由于 $\boldsymbol{A}$, $\boldsymbol{B}$ 的非零元素可以临时产生或放在外存,内存中则不需要存放 $\boldsymbol{A}$, $\boldsymbol{B}$ 的单元。因而,利用这些方法是可以在中小型计算机上解算高阶问题的。不过,这些方法所需迭代次数一般较多,每步计算量也较大,其实际使用的效果还是不够理想的,这一点还有待改进。

附录　代数特征值问题计算程序

一、实对称矩阵的雅可比法程序

$JaC\Theta Bi(A,\ n,\ SV,\ Vec,\ D,\ ts)$

使用说明

本程序是用雅可比法(旋转法)求实对称矩阵的全部特征值和特征向量。所用的计算公式及程序编制中的某些原则可见本章 §10.3 的 10.3.4 节。使用本程序之前，要求将对称矩阵的上三角部分元素以按行排列的顺序存放于一维场 $A[1:n*(n+1)/2]$。本程序工作完毕后，即将求得的特征值和相应特征向量分别存于使用者所指定的场 $D[1:n]$ 和场 $Vec[1:n,\ 1:n]$ 之中，并将旋转变换的总次数存于简变 ts 内。场 A、D、Vec 和简变 ts 必需由使用者在调用本程序之前定义好。如果不需要求特征向量，可令 $SV=0$，程序会自动跳过计算特征向量的步骤，场 Vec 中的内容将不改变。

形式参数表

(1) 输入参数

n——矩阵的阶数。

A——存放对称矩阵上三角部分元素的一维场，其定义为：$A[1:n*(n+1)/2]$。元素按行存放，其次序为：$a_{11},\ a_{12},\ \cdots,\ a_{1n},\ a_{22},\ \cdots,\ a_{2n},\ a_{33},\ \cdots,\ a_{nn}$。

SV——若 SV 为零，表示不需要计算特征向量。否则，程序同时算出特征值和相应特征向量。

(2) 输出参数

D——存放特征值计算结果的一维场，其定义为：$D[1:n]$。特征值是按由大至小的顺序依次存放的，即 $D[1]\geqslant D[2]\geqslant\cdots\geqslant D[n]$。

Vec——存放特征向量计算结果的二维场，其定义为：$Vec[1:n,\ 1:n]$。特征向量的排列次序与 D 中的特征值相对应。其分量排列的方式是先排各特征向量的第一个分量，然后第二个分量，如此等等。

ts——存放旋转变换的总次数。

程序

```
过程 JACθBI(A, N, SV, VEC, D, TS);
  值 N, SV;
  简变 TS;
  场 A, VEC, D;
始  简变 SM, C, S, T, H, G, TA, IZ, TR, G1;
    场 B, Z[1:N];
    若 SV=0 则
            否对于 I=1 到 N 步长 1 执行
```

```
            对于 J=1 到 N 步长 1 执行
            若 I=J 则 1⇒VEC[I, J]
                   否 0⇒VEC[I, J];
0⇒TS; 1⇒IZ;
对于 I=1 到 N 步长 1 执行
  始 0⇒Z[I]; A[IZ]⇒B[I]; B[I]⇒D[I]; IZ+N-I+1⇒IZ
  终;
对于 L=1 到 50 步长 1 执行
SWEP: 始
        0⇒SM; 1⇒IZ;
        对于 I=1 到 N-1 步长 1 执行
          始对于 J=1 到 N-I 步长 1 执行
            §ABS(A[IZ+J])+SM⇒SM; IZ+N+1-I⇒IZ
          终;
  若 SM=0 则转 θUT 否;
  若 L<4 则 0.2*SM/N↑2⇒TR 否 0⇒TR;
    1⇒IZ;
    对于 I=1 到 N-1 步长 1 执行
    始对于 J=I+1 到 N 步长 1 执行
      始 A[IZ+J-I]⇒G1; G1*100⇒G;
        若 4<L 则若 §ABS(D[I])+G=§ABS(D[I])
                   则若 §ABS(D[J])+G=§ABS(D[J])
                     则转 SUN
                     否
                   否
                 否;
        若 §ABS(G1)≤TR 则转 ZUN 否
    RθT: 始 D[I]-D[J]⇒H;
            若 2*§ABS(G1)<§ABS(H)
              则始 2*G1/H⇒T;
            1/§SQRT(1+T↑2)⇒C; T*C⇒S
          终
        否始 0.5*H/G1⇒T;
            §ABS(T)/§SQRT(1+T↑2)⇒C;
            若 T=0 则 1⇒S 否
                      C/T⇒S
          终;
        §SQRT((1+C)/2)⇒C;
        0.5*S/C⇒S;
```

```
          G1*S/C⇒H; S/(1+C)⇒TA;
          Z[I]+H⇒Z[I];
          Z[J]-H⇒Z[J];
          D[I]+H⇒D[I];
          D[J]-H⇒D[J];
          I⇒C; J⇒T;
          对于 K=1 到 N 步长 1 执行
          始 A[C]⇒G; A[T]⇒H; G+S*(H-G*TA)⇒A[C]; H-S*(G+H*TA)⇒
            A[T];
            若 K<I 则 C+N-K⇒C
                  否 C+1⇒C;
            若 K<J 则 T+N-K⇒T
                  否 T+1⇒T
          终;
          若 SV=0 则
                  否
            对于 K=1 到 N 步长 1 执行
            始 VEC[K, I]⇒G; VEC[K, J]⇒H; G+S*(H-G*TA)⇒VEC[K, I];
              H-S*(G+H*TA)⇒VEC[K, J]
            终;
          TS+1⇒TS
        终;
  SUN: 0⇒A[IZ+J-I];
  ZUN:
        终; IZ+N-I+1⇒IZ
      终;
      对于 I=1 到 N 步长 1 执行
        始 B[I]+Z[I]⇒B[I]; B[I]⇒D[I]; 0⇒Z[I]
        终
      终;
θUT:
    对于 I=1 到 N-1 步长 1 执行
      对于 J=I+1 到 N 步长 1 执行
      若 D[I]<D[J] 则
            始 D[I]⇒G; D[J]⇒D[I]; G⇒D[J];
              若 SV=0 则
                否对于 K=1 到 N 步长 1 执行
                  始 VEC[K, I]⇒G; VEC[K, J]⇒VEC[K, I]; G⇒VEC[K, J]
                  终
```

终

否

终;

二、任意实矩阵的广义雅可比法程序

$EIGN(N, A, T, TMx)$

使用说明

本程序是用广义雅可比方法求任意实矩阵的全部特征值和特征向量，其计算公式为：

$$\begin{cases} \boldsymbol{A}^{(1)}=\boldsymbol{A} \\ \boldsymbol{A}^{(k+1)}=\boldsymbol{T}_k^{-1}\cdot\boldsymbol{A}^{(k)}\cdot\boldsymbol{T}_k \\ \boldsymbol{T}_k=\boldsymbol{R}_k\cdot\boldsymbol{S}_k \end{cases} \qquad (k=1,\ 2,\ \cdots)$$

其中 $\boldsymbol{R}_k$ 为旋转矩阵，其元素 $r_{ij}^{(k)}$ 为：

$$r_{ij}^{(k)}=-r_{ji}^{(k)}=-\sin x_k;\quad r_{ii}^{(k)}=r_{jj}^{(k)}=\cos x_k;\quad r_{pq}^{(k)}=\delta_{pq}(\text{其它元素})$$

$\boldsymbol{S}_k$ 为所谓的“剪切”矩阵，其元素 $S_{ij}^{(k)}$ 为：

$$S_{ij}^{(k)}=S_{ji}^{(k)}=-\sinh y_k;\quad S_{ii}^{(k)}=S_{jj}^{(k)}=\cosh y_k;\quad S_{pq}^{(k)}=\delta_{pq}(\text{其它元素})$$

参数 x_k 满足如下方程式：

$$\mathrm{tg}\,2x_k=(a_{ij}^{(k)}+a_{ji}^{(k)})/(a_{ii}^{(k)}-a_{jj}^{(k)})$$

并选择 x_k 使得 $\boldsymbol{A}^{(k+1)}$ 第 i 列的模大于第 j 列者 $(i<j)$。参数 y_k 满足如下方程式：

$$\tanh y_k=(ED-H/2)/(G+2(E^2+D^2))$$

其中

$$E=(a_{ij}^{(k)}-a_{ji}^{(k)})$$

$$D=\cos 2x_k\cdot(a_{ii}^{(k)}-a_{jj}^{(k)})+\sin 2x_k\cdot(a_{ij}^{(k)}+a_{ji}^{(k)})$$

$$G=\sum_{p\neq i,j}(a_{pj}^{(k)2}+a_{jp}^{(k)2}+a_{pi}^{(k)2}+a_{ip}^{(k)2})$$

$$H=\cos 2x_k\Big(2\sum_{p\neq i,j}(a_{ip}^{(k)}\cdot a_{jp}^{(k)}-a_{pi}^{(k)}\cdot a_{pj}^{(k)})\Big)-\sin 2x_k\Big(\sum_{p\neq i,j}(a_{ip}^{(k)2}+a_{pj}^{(k)2}-a_{pi}^{(k)2}-a_{jp}^{(k)2})\Big)$$

矩阵 $\boldsymbol{A}^{(k)}$ 将收敛于分块对角型矩阵，其对角线上子块为 1×1 或如下 2×2 子矩阵：

$$\begin{pmatrix} \tilde{a}_{jj}, & \tilde{a}_{j,j+1} \\ -\tilde{a}_{j,j+1}, & \tilde{a}_{jj} \end{pmatrix}$$

它们分别对应于矩阵 $\boldsymbol{A}$ 的实特征值和复共轭特征值 $\tilde{a}_{jj}\pm i\cdot\tilde{a}_{j,j+1}$。逐次变换矩阵 $\boldsymbol{T}_k$ 之乘积矩阵：

$$\boldsymbol{T}=\boldsymbol{T}_1\cdot\boldsymbol{T}_2\cdots\boldsymbol{T}_k\cdots$$

即为相应特征向量列所构成的矩阵(注意！与复特征值 $\tilde{a}_{ij}\pm i\cdot\tilde{a}_{j,j+1}$ 相对应的特征向量为 $t_j\pm i\cdot t_{j+1}$，t_j 和 t_{j+1} 为矩阵 $\boldsymbol{T}$ 的第 j 列和 $j+1$ 列)。

本程序中，足标对 (i, j) 是按循环方式选取的 $(i<j)$㊀。也有可能 $\boldsymbol{A}^{(k)}$ 收敛于 $a\boldsymbol{I}+\boldsymbol{S}$ 的形式，其中 $\boldsymbol{S}$ 为反对称矩阵。本程序对这种情形不予处理。

本程序对于单构矩阵效果较好，如果矩阵有重的复特征值或为非单构的，收敛可能很慢。为避免浪费过多时间，迭代超过 50 循环后即返回主程序。

㊀ 若非对角元很小，相应变换略去。本程序中是用变量 EP 控制，并取为 10^{-6}。但 EP 之值与所用机器有关，采用其它计算机时，应将此值相应更改。

如果要求复矩阵 $\boldsymbol{A}+i\cdot\boldsymbol{B}$ 的特征值,可以用本程序解 $2n$ 阶实矩阵:

$$\begin{pmatrix} \boldsymbol{A}, & \boldsymbol{B} \\ -\boldsymbol{B}, & \boldsymbol{A} \end{pmatrix}$$

的特征值问题,若原来矩阵的特征值为 $\lambda_1, \lambda_2, \cdots, \lambda_n$, 则 $2n$ 阶矩阵之特征值为 $\lambda_i, \bar{\lambda}_i (i=1, 2, \cdots, n)$。

形式参数表

(1) 输入参数

N——矩阵的阶数。

A——存放矩阵元素的二维场,其定义为: $A[1:n, 1:n]$。矩阵元素按行存放,其次序为: $a_{11}, a_{12}, \cdots, a_{1n}, a_{21}, \cdots, a_{2n}, a_{31}, \cdots, a_{nn}$。

TMX——$TMX=0$, 表示不求特征向量,此时形参中的 T 可代之以与场 A 相应的实在参数,以节省存储量, $tMx<0$ 和 $tMx>0$ 分别表示要求左或右特征向量。

(2) 输出参数

A——本程序工作完毕后,场 A 中存放矩阵 A 的对角线子块(1×1 或 2×2)处,即为矩阵 A 的特征值。1×1 子块对应于实特征值, 2×2 子块对应于复共轭特征值。从左上角至右下角的子块系按相应特征值之模递减的次序排列的。

T——存放求得的特征向量矩阵 T 的二维场,其定义和元素排列的方式与场 A 相同。若矩阵 A 按模递减顺序排列的第 j 个特征值为实数,则矩阵 T 的第 j 行(或列)即为与之相应的左(或右)特征向量。若第 j 个与 $j+1$ 个特征值为共轭复数,则它们相应的左(或右)特征向量为 $t_j \pm i\cdot t_{j+1}$, 其中 t_j 和 t_{j+1} 分别为矩阵 T 的第 j 行(或列)和第 $j+1$ 行(或列)。

TMX——简变 TMX 中记录了总的迭代循环数,若 $0<TMX<50$, 表示迭代过程已收敛。若 $TMX=50$, 表明方法失败。在一个循环内各次变换矩阵均为单位矩阵时, $TMX<0$。

程序

```
过程 EIGN(N, A, T, TMX);
 值 N; 简变 TMX; 场 A, T;
始 简变 EPS, EP, AII, AIJ, AJI, H, G, HJ, AIK, AKI, AIM, AMI, TEP, TEM,
      D,C,E,AKM,AMK,CX,SX,CT2X,SIG,CTX,CS2X,SN2X,TE,TEE,YH,
      DEN,.TNHY, CHY, SHY, C1, C2, S1, S2, TKI,TMI,TIK,TIM,J8,J9,J10;
  1⇒J8; 1⇒J9; 1⇒J10;
  若 TMX=0 则转 BTT 否若 TMX<0 则 −1⇒J9
                                  否 −1⇒J10;
  0⇒T;
  对于 I=1 到 N 步长 1 执行 1⇒T[I, I];
 BTT: 0.000001⇒EP; §SQRT(EP)⇒EPS;
  对于 IT=1 到 50 步长 1 执行
   始若 J8<0 则始 1−IT⇒TMX; 转 DθNE 终否;
    对于 I=1 到 N−1 步长 1 执行
```

```
        始 A[I, I]⇒AII;
          对于 J=I+1 到 N 步长 1 执行
          始 A[I, J]⇒AIJ; A[J, I]⇒AJI;
            若 §ABS(AII+AJI)≤EPS
            则若 §ABS(AIJ-AJI)≤EPS 则
                                                否
              若 §ABS(AII-A[J, J])≤EPS 则
                        否转 BBT
            否转 BBT
          终
        终;
        IT-1⇒TMX; 转 DθNE;
BBT:    -1⇒J8;
        对于 K=1 到 N-1 步长 1 执行
          对于 M=K+1 到 N 步长 1 执行
          始 0⇒H; 0⇒G; 0⇒HJ; 0⇒YH;
            对于 I=1 到 N 步长 1 执行
            始 A[I, K]⇒AIK; A[I, M]⇒AIM; AIK*AIK⇒TE; AIK*AIM⇒TEE;
              YH+TE-TEE⇒YH;
              若 I=K 则否若 I=M 则否
              始 A[K, I]⇒AKI; A[M, I]⇒AMI; H+AKI*AMI-AIK*AIM⇒H;
                TE+AMI*AMI⇒TEP;
                TEE+AKI*AKI⇒TEM; G+TEP+TEM⇒G; HJ-TEP+TEM⇒HJ
              终
            终;
            H+H⇒H; A[K, K]-A[M, M]⇒D; A[K, M]⇒AKM; A[M, K]⇒AMK;
            AKM+AMK⇒C; AKM-AMK⇒E;
            若 §ABS(C)≤EP
            则始 1⇒CX; 0⇒SX 终
            否始 D/C⇒CT2X;
                若 CT2X<0 则 -1⇒SIG
                          否 1⇒SIG;
                CT2X + SIG * §SQRT (1 + CT2X * CT2X) ⇒CTX; SIQ/§SQRT (1+
                  CTX*CTX)⇒SX; SX*CTX⇒CX
              终;
            若 YH<0 则始 CX⇒TEM; SX⇒CX;
                        -TEM⇒SX
                      终否;
            CX*CX-SX * SX ⇒ CS2X; 2*SX*CX ⇒ SN2X; D*CS2X+C*SN2X ⇒ D;
```

```
H*CS2X-HJ*SN2X⇒H; G+2*(E*E+D*D)⇒DEN; (E*D-H/2)/
DEN⇒TNHY;
若 §ABS(TNHY)≤EP
则始 1⇒CHY; 0⇒SHY 终
否始 1/§SQRT(1-TNHY*TNHY)⇒CHY; CHY*TNHY⇒SHY
  终;
CHY*CX-SHY*SX⇒C1; CHY*CX+SHY*SX⇒C2; CHY*SX+SHY*
CX⇒S1; -CHY*SX+SHY*CX⇒S2;
若 §ABS(S1)≤EP
则若 §ABS(S2)≤EP 则转 AAT 否否;
1⇒J8;
对于 I=1 到 N 步长 1 执行
始 A[K, I]⇒AKI; A[M, I]⇒AMI;
    C1*AKI+S1*AMI⇒A[K, I];
    S2*AKI+C2*AMI⇒A[M, I];
    若 J9<0 则始 T[K, I]⇒TKI;
           T[M, I]⇒TMI;
           C1*TKI+S1*TMI⇒T[K, I];
           S2*TKI+C2*TMI⇒T[M, I]
             终
           否
终;
对于 I=1 到 N 步长 1 执行
始 A[I, K]⇒AIK; A[I, M]⇒AIM;
  C2*AIK-S2*AIM⇒A[I, K];
  -S1*AIK+C1*AIM⇒A[I, M];
  若 J10<0 则
  始 T[I, K]⇒TIK; T[I, M]⇒TIM;
      C2*TIK-S2*TIM⇒T[I, K];
      -S1*TIK+C1*TIM⇒T[I, M]
  终
           否
终;
AAT:
   终
终;
50⇒TMX;
DθNE:
终;
```

三、化实对称矩阵为三对角型程序
$HTRD(A, n, EPS, C, B)$
$HBAK(A, n, n1, n2, z)$

使用说明

本程序是用镜像映射矩阵作相似变换把给定实对称矩阵 $\boldsymbol{A}$ 化为三对角型矩阵 $\boldsymbol{A}_{n-1}$。计算公式及程序安排可参见本章§10.3的(10.3.43)式及有关的讨论。程序中包括两个过程,过程 $HTRD$ 是将矩阵 $\boldsymbol{A}$ 化为三对角型 $\boldsymbol{A}_{n-1}$, 过程 $HBAK$ 是把用其它程序求得的 $\boldsymbol{A}_{n-1}$ 的特征向量转换成矩阵 $\boldsymbol{A}$ 的特征向量。两个过程需要分别进行调用。

形式参数表

(1) 过程 $HTRD(A, n, EPS, C, B)$;

输入参数

n——矩阵的阶数。

A——存放矩阵 A 上三角部分元素的一维场,其定义为: $A[1:n*(n+1)/2]$。元素按行存放,其次序为: $a_{11}, a_{12}, \cdots, a_{1n}, a_{22}, \cdots, a_{2n}, a_{33}, \cdots, a_{nn}$。

EPS——控制某次镜像映射变换可否忽略的误差界,一般可以取为计算机能够表示的最小正数。如果尾数部分的字长 t 远比计算机中最大允许阶码 q 小得多,也可令 $EPS=2^{-(q-t)}$。

输出参数

C——存放求得的三对角矩阵 A_{n-1} 对角线元素的一维场,其定义为: $C[1:n]$。对角线元素存放顺序为 $C_1, C_2, \cdots, C_n$。

B——存放求得的三对角矩阵 A_{n-1} 次对角线元素的一维场,其定义为: $B[1:n]$。存放方式为 $B[1]$ 任意,从 $B[2]$ 起依次存放 $b_2, b_3, \cdots, b_n$。

(2) 过程 $HBAK(A, n, n1, n2, z)$;

输入参数

n——矩阵的阶数。

A——即过程 $HTRD$ 中的场 A, 该过程执行完毕后,场 A 中存放着逐次变换矩阵 Hi, 本过程中将要用到这些矩阵。

$n1$——要求转换的第一个特征向量的编号(或场 z 第二个下标的下界)。

$n2$——要求转换的最后一个特征向量的编号(或场 z 第二个下标的上界)。

z——存放要求进行转换的特征向量的二维场,其定义为: $z[1:n, n1:n2]$。存放方式为先依次存各向量的第一个分量,然后第二个分量,如此等等。

输出参数

z——本程序工作完毕后,场 z 中按前述同样方式存放着要求的矩阵 $\boldsymbol{A}$ 的第 $n1$ 至 $n2$ 个特征向量。

程序

```
过程 HTRD(A, N, EPS, C, B);
值 N, EPS;
```

```
场 A, C, B;
始  简变 IZ, JK, F, G, H, GG, HH;
   对于 I=1 到 N-1 步长 1 执行
   始 0⇒H; 1+N*(I-1)-(I-1)*(I-2)/2⇒IZ;
      对于 K=N 到 I+1 步长 -1 执行
        始 A[IZ+K-I]⇒F; F⇒C[K];
          F*F+H⇒H
        终;
      若 H≤EPS 则始 0⇒H; 0⇒GG;
                  转 SKIP 终
                否;
     若 F<0 则 1⇒G 否 -1⇒G;
     G*§SQRT(H)⇒GG;
     H-F*GG⇒H; F-GG⇒C[I+1];
     0⇒F; C[I+1]⇒A[IZ+1];
     对于 J=I+1 到 N 步长 1 执行
     始 0⇒G; IZ+N+J-2*I⇒JK;
        对于 K=I+1 到 N 步长 1 执行
          始 G+A[JK]*C[K]⇒G;
            若 K<J 则 JK+N-K⇒JK
                 否 JK+1⇒JK
          终;
          G/H⇒G; G⇒B[J];
          F+G*C[J]⇒F
     终;
     0.5*F/H⇒HH; N*(N+1)/2⇒JK;
     对于 J=N 到 I+1 步长 -1 执行
       始 C[J]⇒F; B[J]-HH*F⇒G; G⇒B[J];
        对于 K=N 到 J 步长 -1 执行
          始 A[JK]-F*B[K]-G*C[K]⇒A[JK]; JK-1⇒JK
          终
       终;
     SKIP: A[IZ]⇒C[I]; H⇒A[IZ]; GG⇒B[I+1]
     终;
  A[JK+1]⇒C[M]
终;
过程 HBAK(A, N, N1, N2, Z);
  值 N, N1, N2;
  场 A, Z;
```

```
始 简变 IZ, H, S;
   对于 I=N-1 到 1 步长 -1 执行
     始 1+N*(I-1)-(I-1)*(I-2)/2⇒IZ;
       A[IZ]⇒H;
       若 H=0 则
              否对于 J=N1 到 N2 步长 1 执行
                  始 0⇒S;
                    对于 K=N 到 I+1 步长 -1 执行
                    S+A[IZ+K-I]*Z[K, J]⇒S; S/H⇒S;
                    对于 K=N 到 I+1 步长 -1 执行
                      Z[K, J]-S*A[IZ+K-I]⇒Z[K, J]
                  终
     终
终;
```

四、对称三对角型矩阵的区间分半法程序

FeNbAN(*C*, *B*, *x*, *n*, *n*1, *n*2, mins, *EPS*1, *EPS*2, *H*)

使用说明

本程序是用区间分半法计算对称三对角型矩阵的按由小至大次序排列的第 $n1$ 个到第 $n2$ 个特征值。假设考虑的对称三对角型矩阵形如(10.3.44)的 $\boldsymbol{A}_{n-1}$。程序要求使用者将 $\boldsymbol{A}_{n-1}$ 的对角线元素依次存放于场 $C[1:n]$，次对角线元素存于场 $B[1:n]$ ($B[1]$ 可为任意值)。并要求给出控制分半过程结束的允许误差 $EPS1$ 和代替等于零的 $q_i(\lambda)$ 的小量 $\min s$。一般可令 $EPS1=2^{-t}\cdot\max_i\{|b_{i-1}|+|C_i|+|b_i|\}$；$\min s=2^{-t}$。如果要求绝对值较小的特征值尽可能准确，也可令 $EPS1$ 为其允许误差。程序求得的结果是矩阵 $\boldsymbol{A}_{n-1}$ 的按由小至大顺序排列的第 $n1$ 至 $n2$ 个特征值，并将其放到场 $X[n1:n2]$ 中。程序也同时给出总的迭代次数(分半次数)和求得的特征值的误差界，并分别存于简变 H 和 $EPS2$。一般来说，$EPS2$ 是远大于所得特征值的真正误差的，仅能作为参考。

形式参数表

(1) 输入参数

n——矩阵的阶数。

$n1$——要求的第一个特征值的编号(假定矩阵 A_{n-1} 的特征值按由小到大的次序排列)。

$n2$——要求的最后一个特征值的编号。

$\min s$——所用计算机上，$1+\varepsilon\neq1$ 的最小正数 ε。如果计算机的字长(指尾数部分而言)为 t，一般可令 $\min s=2^{-t}$。

$EPS1$——控制分半过程结束的允许误差。一般可令其为 $2^{-t}\cdot\max_i\{|b_{i-1}|+|C_i|+|b_i|\}$，或视使用者需要而定。

C——存放矩阵 $\boldsymbol{A}_{n-1}$ 的对角线元素的一维场，其定义为：$C[1:n]$。存放次序为：

$C_1, C_2, \cdots, C_n$。

B——存放矩阵 A_{n-1} 的次对角线元素的一维场，其定义为：$B[1:n]$。存放方式是 $B[1]$ 可为任意值，从 $B[2]$ 起，依次存放 $b_2, b_3, \cdots, b_n$。

(2) 输出参数

X——存放特征值计算结果的一维场，其定义为：$X[n1:n2]$。存放次序是由小到大的次序，即 $X[n1] \leqslant X[n1+1] \leqslant \cdots \leqslant X[n2]$。

H——存放求出所要的全部特征值的总迭代次数。

$EPS2$——计算所得诸特征值的误差界。

程序

```
过程 FENBAN(C, B, X, N, N1, N2, MINS, EPS1, EPS2, H);
  值 N, N1, N2, MINS;
  场 C, B, X;
  简变 EPS1, EPS2, H;
始  简变 Z, XMIN, XMAX;
    0⇒B[1]; C[N]-§ABS(B[N])⇒XMIN; C[N]+§ABS(B[N])⇒XMAX;
    对于 I=N-1 到 1 步长 -1 执行
      始 §ABS(B[I])+§ABS(B[I+1])⇒Z;
        若 XMAX<C[I]+Z 则 C[I]+Z⇒XMAX 否;
        若 C[I]-Z<XMIN 则 C[I]-Z⇒XMIN 否
      终;
    若 0<XMAX+XMIN 则 XMAX⇒Z 否
                              -XMIN⇒Z;
    Z*MINS⇒EPS2;
    若 EPS1≤0 则 EPS2⇒EPS1 否;
    7*EPS2+EPS1/2⇒EPS2;
  始  简变 A, Q, Q1, XU, XS;
      场  WU[N1:N2];
        XMAX⇒XS;
      对于 I=N1 到 N2 步长 1 执行
      始 XMAX⇒X[I]; XMIN⇒WU[I] 终;
      0⇒H;
      对于 K=N2 到 N1 步长 -1 执行
        始 XMIN⇒XU;
          对于 I=K 到 N1 步长 -1 执行
            若 XU<WU[I]则始 WU[I]⇒XU;
                            转 CθNT
                          终
                        否;
       CθNT: 若 X[K]<XS 则 X[K]⇒XS 否;
```

```
        对于 X1=(XU+XS)/2
          当 2*MINS*(§ABS(XU)+§ABS(XS))+EPS1<XS-XU
          执行
         始 H+1⇒H;
          0⇒A; 1⇒Q;
          对于 I=1 到 N 步长 1 执行
            始若 Q=0
              则 §ABS(B[I])/MINS⇒Q1
              否 B[I]*B[I]/Q⇒Q1; C[I]-X1-Q1⇒Q;
              若 Q<0 则 A+1⇒A 否
            终;
          若 A<K 则
            若 A<N1 则始 X1⇒XU;
                         X1⇒WU[N1]
                       终
                   否始 X1⇒XU; X1⇒WU[A+1];
                        若 X1<X[A]
                          则 X1⇒X[A]
                          否
                       终
              否 X1⇒XS
          终;
          (XS+XU)/2⇒X[K]
   终
  终
终;
```

五、求对称三对角型矩阵特征向量的反幂法程序 $INVER(C, B, X, n, n1, n2, VEC, \min s)$

使用说明

本程序是用反幂法计算对称三对角型矩阵 $\boldsymbol{A}_{n-1}$(见(10.3.44)式)相应于已知近似特征值的特征向量。

程序假定特征值已由区间分半过程或其它过程求得(存于场 $X[n1:n2]$), 本程序求得的相应特征向量存于场 $VEC[1:n, n1:n2]$。参数 $\min s$ 与区间分半法程序中相同。在进行计算时,首先将 $\boldsymbol{A}_{n-1}-\lambda_i I$ 按列主元素消去法进行分解。所得的上三角型矩阵 $\boldsymbol{U}$ 的主对角线元素,次对角线元素及第三对角线元素分别存于场 D, E, $F[1:n]$, 所得的单位下三角型矩阵 $\boldsymbol{L}$ 的次对角线元素存于场 $Y[2:n]$, 行交换信息存于场 $Int[2:n]$。然后再进行迭代。通常求出一个向量的迭代次数不超过 5, 否则认为该向量求不出来,而将相应的迭代次数记数

单元送以 6 作为失败的标志。如此逐个向量进行迭代，直到求完所有向量为止。

形式参数表

(1) 输入参数

n——矩阵的阶数。

$n1$——要求的第一个特征向量相应的编号(即其相应的特征值按由小至大次序的编号)。

$n2$——要求的最后一个特征向量相应的编号。

min s——所用计算机上 $1+\varepsilon\neq1$ 的最小正数 ε。若计算机的字长为 t(指尾数部分而言)，一般可令 $\min s\cong2^{-t}$。

X——存放特征值近似值的一维场，其定义为：$X[n1:n2]$。其中存放用区间分半法或其它过程求得的矩阵 A_{n-1} 的第 $n1$ 个至 $n2$ 个特征值(按由小至大次序排列)。

C——存放矩阵 A_{n-1} 对角线元素的一维场，其定义为：$C[1:n]$。存放次序为 C_1, C_2, $\cdots$, C_n。

B——存放矩阵 A_{n-1} 次对角线元素的一维场，其定义为：$B[1:n]$。存放方式为 $B[1]$ 任意，从 $B[2]$ 起依次存放 b_2, b_3, $\cdots$, b_n。

(2) 输出参数

Vec——存放反幂法求得的特征向量的二维场，其定义为：$Vec[1:n, n1:n2]$。存放方式是将 $n2-n1+1$ 个特征向量视为一个 $n\times(n2-n_1+1)$ 维矩阵，其每一列为一个特征向量，然后将这个矩阵的元素以按行排列的次序存于场 Vec 中。

程序

```
过程 INVER(C, B, X, N, N1, N2, VEC, MINS);
 值 N, N1, N2, MINS;
 场 C, B, X, VEC;
始  简变 U, V, XU, BI, S, X1, X2, NØRM, EPS, EPS1, EPS2, GP, TS;
    场 Z, D, E, F[1:N], INT, Y[2:N], CTS[N1:N2];
    §ABS(C[1])⇒NØRM;
    对于 I=2 到 N 步长 1 执行
    NØRM+§ABS(C[I])+§ABS(B[I])⇒NØRM;
    -NØRM⇒X1;
    NØRM*10↑(-4)⇒EPS;
    NØRM*MINS⇒EPS1;
    EPS1*N⇒EPS2; 0⇒GP; 1⇒S;
    对于 K=N1 到 N2 步长 1 执行
    始 1⇒TS; X[K]⇒X2;
      若 X2-X1<EPS 则 GP+1⇒GP 否 0⇒GP;
      若 X2≤X1 则 X1+EPS1⇒X2 否;
      EPS2/§SQRT(N)⇒U;
      对于 I=1 到 N 步长 1 执行 U⇒Z[I];
      C[1]-X2⇒U; B[2]⇒V;
```

```
对于 I=2 到 N 步长 1 执行
始 B[I]⇒BI;
  若 §ABS(U)≤§ABS(BI) 则 1⇒INT[I]
                         否 0⇒INT[I];
  若 INT[I]=1 则
      始 U/BI⇒XU; XU⇒Y[I];
        BI⇒D[I-1]; C[I]-X2⇒E[I-1];
        若 I=N 则 0⇒F[I-1]
               否 B[I+1]⇒F[I-1];
        V-XU*E[I-1]⇒U; -XU*F[I-1]⇒V
      终    否
      始 BI/U⇒XU; XU⇒Y[I]; U⇒D[I-1]; V⇒E[I-1]; 0⇒F[I-1];
        C[I]-X2-XU*V⇒U;
        若 I=N 则否 B[I+1]⇒V
      终;
  若 U=0 则 EPS1⇒D[N]
         否 U⇒D[N];
  0⇒E[N]; 0⇒F[N];
终;
NEWZ: 对于 I=N 到 1 步长 -1 执行
    始(Z[I]-U*E[I]-V*F[I])/D[I]⇒Z[I]; U⇒V; Z[I]⇒U
    终;
    对于 J=K-GP 到 K-1 步长 1 执行
    始 0⇒XU;
      对于 I=1 到 N 步长 1 执行
      XU+Z[I]*VEC[I, J]⇒XU;
      对于 I=1 到 N 步长 1 执行
      Z[I]-XU*VEC[I, J]⇒Z[I]
    终;
    0⇒NΘRM;
    对于 I=1 到 N 步长 1 执行
    NΘRM+§ABS(Z[I])⇒NΘRM;
    若 NΘRM<1 则
    始若 TS=5 则始 6⇒CTS[K];
                   转 END
               终
             否;
      若 NΘRM=0 则始 EPS2⇒Z[S];
                  若 S=N 则 1⇒S
```

```
                        否 S+1⇒S
                    终
                  否始 EPS2/NΘRM⇒XU;
                      对于 I=1 到 N 步长 1 执行
                      Z[I]*XU⇒Z[I]
                    终;
          对于 I=2 到 N 步长 1 执行
            若 INT[I]=1 则
              始 Z[I-1]⇒U; Z[I]⇒Z[I-1]; U-Y[I]*Z[I]⇒Z[I]
              终
                      否
              Z[I]-Y[I]*Z[I-1]⇒Z[I];
          TS+1⇒TS;
          转 NEWZ
        终
                否;
        0⇒U;
          对于 I=1 到 N 步长 1 执行
          U+Z[I]↑2⇒U; 1/§SQRT(U)⇒XU;
          对于 I=1 到 N 步长 1 执行
          XU*Z[I]⇒VEC[I, K]; TS⇒CTS[K];
      END: X2⇒X1
    终
  终;
```

六、化带型实对称矩阵为三对角型程序 DAIRD(A, n, m, SV, V, D, E)

使用说明

本程序是用本章§10.3表10.3所示的一系列旋转变换将带型实对称矩阵 $\boldsymbol{A}$ 化为三对角型矩阵 $\widetilde{\boldsymbol{D}}$:

$$\widetilde{D}=\begin{bmatrix} C_1 & b_2 & & & \\ b_2 & C_2 & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & b_n \\ & & & b_n & C_n \end{bmatrix}$$

根据使用者的要求，也可同时算出逐次旋转矩阵的乘积矩阵 $\boldsymbol{V}(\boldsymbol{D}=\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V})$。

为了节省存储单元，程序要求使用者将矩阵 $\boldsymbol{A}$ 上三角部分的非零元素带，以按行排列的次序存于二维场 $A[1:n,\ 0:m]$ 内。例如：

$$\boldsymbol{A}=\begin{bmatrix} 6 & -4 & 1 & & & \\ -4 & 6 & -4 & 1 & \mathbf{0} & \\ 1 & -4 & 6 & -4 & 1 & \\ & 1 & -4 & 6 & -4 & 1 \\ & & 1 & -4 & 6 & -4 \\ & \mathbf{0} & & 1 & -4 & 6 \end{bmatrix}$$

则场 A 为：

6	−4	1
6	−4	1
6	−4	1
6	−4	1
6	−4	×
6	×	×

其中三个“×”元素可为任意值。

程序还假定矩阵阶数 n 和上三角部分的非零对角线条数 m 以及是否需要计算矩阵 $\boldsymbol{V}$ 的信息 SV，均由使用者给出。程序将求得的三对角型矩阵 $\breve{\boldsymbol{D}}$ 之对角线元素 C_i 存于场 $D[1:n]$ 中，次对角线元素 b_i 存于场 $E[1:n]$ ($E[1]$ 可为任意数) 中。我们可将其作为初始数据直接引用区间分半法程序。此外，若 $\boldsymbol{SV}\neq 0$，程序将计算矩阵 $\boldsymbol{V}$，并把它存于场 $V[1:n,\ 1:n]$ 中，作为用反幂法程序求得矩阵 $\widetilde{D}$ 的特征向量 $\boldsymbol{Z}_i$ 后，计算 $\boldsymbol{X}_i=\boldsymbol{V}\cdot\boldsymbol{Z}_i$ 之用。

形式参数表

(1) 输入参数

n——矩阵的阶数。

m——矩阵的“半带宽”(总带宽为 $2m+1$)。

A——存放带型对称矩阵上三角部分非零元素的二维场，其定义为：$A[1:n,\ 0:m]$。元素按行存放，其次序为：$a_{11},\ a_{12},\ \cdots,\ a_{1,m+1},\ a_{22},\ \cdots,\ a_{2,m+2},\ a_{33},\ \cdots,\ a_{nn}$ (参看前面的例)。

SV——若 $SV=0$，程序不计算逐次旋转矩阵的乘积，否则，程序将逐次旋转矩阵连乘起来，并以按行排列的次序存于场 V。

(2) 输出参数

D——存放求得的三对角型矩阵 $\widetilde{D}$ 对角线元素的一维场，其定义为：$D[1:n]$。对角线元存放的次序为 $C_1,\ C_2,\ \cdots,\ C_n$。

E——存放求得的三对角矩阵 $\widetilde{D}$ 次对角线元素的一维场，其定义为：$E[1:n]$。存放方式是 $E[1]$ 可为任意值，从 $E[2]$ 起依次存放 $b_2,\ \cdots,\ b_n$。

V——存放逐次旋转矩阵乘积的二维场,其定义为: $V[1:n, 1:n]$, 乘积矩阵的元素按行存放。

程序

```
过程 DAIRD(A, N, M, SV, V, D, E);
  值 N, M, SV;
  场 A, V, D, E;
始    简变 C, S, C2, S2, CS, CT, U, U1, G, MAXJ, MAXL, UGL;
      若 SV=0 则
              否对于 I=1 到 N 步长 1 执行
                对于 J=1 到 N 步长 1 执行
                  若 I=J 则 1⇒V[I, J]
                        否 0⇒V[I, J];
      对于 I=1 到 N-1 步长 1 执行
        始若 N-I<M 则 N-I⇒MAXJ 否 M⇒MAXJ;
          对于 J=MAXJ 到 2 步长 -1 执行
            始对于 K=I+J 到 N 步长 M 执行
              始若 K=I+J 则始若 A[I, J]=0 则
                                转 ENDJ 否;
                          A[I, J-1]/A[I, J]⇒CT; I⇒UGL
                        终
                      否始若 G=0
                          则转 ENDJ
                          否;
                          A[K-M-1, M]/G⇒CT; K-M⇒UGL
                        终;
              1/§SQRT(1+CT↑2)⇒S;
              CT*S⇒C; C*C⇒C2;
              S*S⇒S2; C*S⇒CS;
              C2*A[K-1, 0]+2*CS*A[K-1, 1]+S2*A[K, 0]⇒U;
              S2*A[K-1, 0]-2*CS*A[K-1, 1]+C2*A[K, 0]⇒U1;
              (A[K, 0]-A[K-1, 0])*CS+A[K-1, 1]*(C2-S2)⇒A[K-1, 1];
              U⇒A[K-1, 0];
              U1⇒A[K, 0];
              若 K=I+J 则
                      否
                    C*A[K-M-1, M]+S*G⇒A[K-M-1, M];
              对于 L=UGL 到 K-2 步长 1 执行
              始 C*A[L, K-L-1]+S*A[L, K-L]⇒U; C*A[L, K-L]
                 -S*A[L, K-L-1]⇒A[L, K-L]; U⇒A[L, K-L-1]
```

```
        终;
      若 N-K<M-1 则 N-K⇒MAXL
                    否 M-1⇒MAXL;
      对于 L=1 到 MAXL 步长 1 执行
        始 C*A[K-1, L+1]+S*A[K, L]⇒U;
           C*A[K, L]-S*A[K-1,L+1]⇒A[K, L];
           U⇒A[K-1, L+1]
        终;
      若 K+M≤N 则始 S*A[K, M]⇒G;
                    C*A[K, M]⇒A[K, M]
                  终
                否;
      若 SV=0 则
              否
                对于 L=1 到 N 步长 1 执行
                  始 C*V[L, K-1]+S*V[L, K]⇒U; C*V[L, K]-S*V[L,
                     K-1]⇒V[L, K]; U⇒V[L, K-1]
                  终
        终;
     ENDJ:
     终;
     A[I, 0]⇒D[I]; A[I, 1]⇒E[I+1]
   终;
   A[N, 0]⇒D[N]
终;
```

七、化 $Ax=\lambda Bx$ 为普通特征值问题程序

ABRD(A, B, n, t)

ABRE$(B, n, n1, n2, z)$

使用说明

本程序是按(10.3.51)式将广义特征值问题 $\boldsymbol{Ax}=\lambda\boldsymbol{Bx}$ 化为普通的对称矩阵特征值问题。程序假定使用者将对称矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的上三角部分元素以按行排列的顺序存放于场 A, $B[1:n*(n+1)/2]$。程序的第一个过程 ABRD(A, B, n, t)是求出 $\boldsymbol{L}^{-1}\cdot\boldsymbol{A}\cdot\boldsymbol{L}^{-T}$ 和 $\boldsymbol{L}^T$ 的上三角部分元素并按行排列地存于场 A 和 B。第二个过程 ABRE$(B, n, n1, n2, z)$是对已求得的矩阵 $\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{L}^{-T}$ 的特征向量 $\boldsymbol{Z}_i$ 左乘以 L^{-T}，从而转换成要求的特征向量 $\boldsymbol{x}_i$。

程序对于矩阵 $\boldsymbol{B}$ 在舍入误差范围内是否正定能自动作出判断。发现非正定时，则停止计算。此外，若矩阵 $\boldsymbol{B}$ 事先已进行过分解并已将 $\boldsymbol{L}^T$ 按行存于场 B，则可令参数 T 为负值，程序执行过程中会自动跳过分解矩阵 $\boldsymbol{B}$ 的步骤。

形式参数表

(1) 过程 ABRD(A, B, n, t)：

输入参数

n——矩阵的阶数。

t——若 $t<0$，表示使用本程序前已将矩阵 B 分解，并将其上三角因子 L^T 按行存于场 B，存放方式与矩阵 B 的元素相同。

A——存放矩阵 A 上三角部分元素的一维场，其定义为：$A[1:n*(n+1)/2]$。元素按行存放，其次序为 a_{11}, a_{12}, a_{13}, $\cdots$, a_{1n}, a_{22}, $\cdots$, a_{2n}, a_{33}, $\cdots$, a_{nn}。本程序工作完毕后，场 A 中按上述同样方式存放着矩阵 $L^{-1}AL^{-T}$ 的上三角部分元素。

B——存放矩阵 B 上三角部分元素的一维场，其定义与存放方式与场 A 相同。本程序工作完毕后，场 B 中按同样方式存放着矩阵 L^T 的上三角部分元素。

(2) 过程 ABRE(B, n, $n1$, $n2$, z)：

输入参数

n——矩阵的阶数。

$n1$——需要转换的第一个特征向量的编号(或在场 Z 中的列下标)。

$n2$——需要转换的最后一个特征向量的编号(或在场 Z 中的列下标)。

B——存放矩阵 L^T 上三角部分元素的一维场，其定义与存放方式与过程 ABRD 内相同。

Z——按行存放 $n\times(s2-s1+1)$ 维矩阵 z 的二维场，其定义为：$Z[1:n, s1:s2]$，其中 $1\leqslant s1\leqslant n1$, $n2\leqslant s2\leqslant n$。矩阵 Z 的第 $n1$ 列至 $n2$ 列为需要转换的特征向量。本程序工作完毕后，矩阵 Z 的第 $n1$ 列至 $n2$ 列(即场 z 中第二个下标为 $n1$ 至 $n2$ 的 $n*(n2-n1+1)$ 个单元)为要求的特征向量 x_i。

程序

```
过程 ABRD(A, B, N, T);
  值 N; 简变 T;
  场 A, B;
始  简变 X, Y  JK;
   若 T<0 则
         否对于 I=1 到 N 步长 1 执行
            对于 J=I 到 N 步长 1 执行
            始 0⇒JK; 0⇒X;
              对于 K=1 到 I-1 步长 1 执行
              始 X+B[I+JK]*B[J+JK]⇒X; JK+N-K⇒JK
              终;
              B[J+JK]-X⇒X;
              若 I=J 则若 X⩽0 则
                始 印+X, I;
                  停 666
```

```
                终
                    否 §SQRT(X)⇒Y
                    否;
                X/Y⇒B[J+JK]
              终;
  对于 I=1 到 N 步长 1 执行
    对于 J=1 到 I 步长 1 执行
      始 0⇒X; 0⇒JK;
        对于 K=1 到 J-1 步长 1 执行
          始 X+A[I+JK]*B[J+JK]⇒X;
             JK+N-K⇒JK
          终;
          (A[I+JK]-X)/B[J+JK]⇒A[I+JK]
      终;
  对于 I=1 到 N 步长 1 执行
    对于 J=I 到 N 步长 1 执行
      始 0⇒X; 0⇒Y; 0⇒JK;
        对于 K=1 到 J-1 步长 1 执行
          始 X+A[I+Y]*B[J+JK]⇒X;
             JK+N-K⇒JK;
             若 K<I 则 Y+N-K⇒Y
                   否 Y+1⇒Y
          终;
        (A[I+Y]-X)/B[J+JK]⇒A[I+Y]
      终
终;
过程 ABRE(B, N, N1, N2, Z);
  值 N, N1, N2;
  场 B, Z;
始  简变 X, IZ;
    对于 J=N1 到 N2 步长 1 执行
      对于 I=N 到 1 步长 -1 执行
        始 1+(I-1)*N-(I-1)*(I-2)/2⇒IZ;
           Z[I, J]⇒X;
          对于 K=I+1 到 N 步长 1 执行
            X-B[IZ+K-I]*Z[K, J]⇒X;
            X/B[IZ]⇒Z[I, J]
        终
终;
```

八、QR 方法求任意实矩阵全部特征值程序

QRMT(*n*, *EPS*, *EtA*, *A*, *RE*, *IM*, *TS*)

使用说明

本程序是按本章 10.3.3 节所述的 QR 方法求任意实矩阵的全部特征值。详细的计算步骤及公式，读者可参阅 §10.3.3 的最后一段。为简单起见，程序中略去了初始矩阵平衡的步骤。

整个程序分为两部分，第一部分是用镜像映射矩阵将原始矩阵 $\boldsymbol{A}$ 化为上海森堡型，并将其仍旧存于矩阵 $\boldsymbol{A}$ 的原来位置上，其计算公式为：

$$\begin{cases} \boldsymbol{A}_1=\boldsymbol{A} \\ \boldsymbol{U}_r^T=(0,\ 0,\ \cdots,\ 0,\ a_{r+1,r}^{(r)}+\mathrm{sign}(a_{r+1,r}^{(r)})\sigma_r^{1/2},\ a_{r+2,r}^{(r)},\ \cdots,\ a_{n,r}^{(r)})^T \\ \sigma_r=(a_{r+1,r}^{(r)})^2+(a_{r+2,r}^{(r)})^2+\cdots+(a_{n,r}^{(r)})^2 \\ H_r=(\sigma_r^{1/2}+|a_{r+1,r}^{(r)}|)\sigma_r^{1/2} \\ \boldsymbol{B}_{r+1}=\boldsymbol{A}_r-\boldsymbol{U}_r[(\boldsymbol{U}_r^T\cdot\boldsymbol{A}_r)/\boldsymbol{H}_r] \\ \boldsymbol{A}_{r+1}=\boldsymbol{B}_{r+1}-[(\boldsymbol{B}_{r+1}\cdot\boldsymbol{U}_r)/\boldsymbol{H}_r]\cdot\boldsymbol{U}_r^T \\ (r=1,\ 2,\ \cdots,\ n-2) \end{cases}$$

如果某个 σ_r 很小，由于计算误差可能导致相应的变换矩阵正交性差，故将该次变换略去，并认为 $\sigma_r=0$。这样作并不影响所得特征值的精度。本程序要求使用者给出一个控制常数 EtA 来决定 σ_r 是否应代之以零，一般来说，EtA 可取为略大于计算机上所能表示的最小正数。

程序的第二部分是用 QR 方法求上海森堡型矩阵的特征值。每进行一次双步 QR 变换称为迭代一次。每次迭代前，按如下公式检查一下是否有单个小的次对角线元素可以忽略：

$$|a_{l,l-1}|\leqslant \mathrm{EPS}*(|a_{l-1,l-1}|+|a_{ll}|)$$

其中 EPS 为计算机上 $1+\mathrm{EPS}\neq 1$ 的最小正数，一般可令 $\mathrm{EPS}=2^{-t}$，t 为计算机的字长(指尾数部分而言)。然后，按如下公式计算相应于第 m 个对角线元的第一次变换矩阵的三个非零元素：

$$\begin{cases} p_m=[(a_{nn}-a_{mm})\cdot(a_{n-1,n-1}-a_{mm})-a_{nn-1}\cdot a_{n-1,n}]/a_{m+1,m}+a_{m,m+1} \\ q_m=a_{m+1,m+1}-a_{mm}-(a_{nn}-a_{mm})-(a_{n-1,n-1}-a_{mm}) \\ r_m=a_{m+2,m+1} \end{cases}$$

如果有：

$$|a_{m,m-1}|(|q_m|+|r_m|)\leqslant \mathrm{EPS}*|p_m|*\left(\sum_{i=m-1}^{m+1}|a_{ii}|\right)$$

则表明经过相应变换后 $(m+1,\ m-1)$ 与 $(m+2,\ m-1)$ 两个元素可以忽略不计，整个矩阵的第 m 行以前仍为上海森堡型，那么变换即可从第 m 列开始。此外，若某个特征值已迭代 10 次(或 20 次)仍不收敛，程序即按如下公式进行一次特殊移位：

$$k_1+k_2=1.5(|a_{n,n-1}|+|a_{n-1,n-2}|)$$

$$k_1\cdot k_2=(|a_{n,n-1}|+|a_{n-1,n-2}|)^2$$

如果迭代 30 次后仍不收敛，即认为方法失败，程序自动停机(停机号码为 888)。若继续启动，则返回主程序。

形式参数表

(1) 输入参数

n——矩阵的阶数。

EPS——计算机上 $1+\varepsilon\neq 1$ 的最小正数 ε。一般可取为 2^{-t}(t 为计算机尾数部分的字长)。

EtA——决定 σ_r 是否可以忽略的控制常数，一般取其略大于计算机上能够表示的最小正数，若阶码允许变化范围 q 比 t 大得多，也可取 $EtA\sim 2^{-(q-t)}$。

A——存放原始矩阵元素的二维场，其定义为：$A[1:n, 1:n]$。矩阵元素按行存放，其次序为：$a_{11}, a_{12}, \cdots, a_{1n}, a_{21}, a_{22}, \cdots, a_{2n}, a_{31}, \cdots, a_{nn}$。程序工作完毕后，场 A 中的内容即已破坏。

(2) 输出参数

RE——存放求得的特征值实部的一维场，其定义为 $RE[1:n]$。

IM——存放求得的特征值虚部的一维场，其定义为：$IM[1:n]$。

TS——存放求得每个特征值所需迭代次数的一维场，其定义为：$TS[1:n]$。如果特征值为一对共轭复根，它们是同时求得的，则第一个的迭代次数与第二个反号。

程序

```
过程 QRMT(N, EPS, ETA, A, RE, IM, TS);
 值 N, EPS, ETA; 场 A, RE, IM, TS;
始  简变P, Q, R, S, T, W, X, Y, Z, ITS, N1, LL, MM, NN, JJ;
   N-1⇒N1;
   对于 M=2 到 N-1 步长 1 执行
   始 0⇒X;
     对于 I=N 到 M 步长 -1 执行
     始 A[I, M-1]⇒P; P⇒RE[I];
       若 §ABS(P)≤X 则否 §ABS(P)⇒X
     终;
     若 X≤ETA 则始 0⇒Q; 转 SKIP 终否;
     0⇒S;
     对于 I=N 到 M 步长 -1 执行
     始 RE[I]/X⇒P; P⇒RE[I]; S+P*P⇒S
     终;
               若 0≤P 则 -1⇒T 否 1⇒T;
     T*§SQRT(S)⇒Q; S-P*Q⇒S; P-Q⇒RE[M]; Q*X⇒Q;
     对于 J=M 到 N 步长 1 执行
     始 0⇒P;
       对于 I=N 到 M 步长 -1 执行
```

```
        P+RE[I]*A[I, J]⇒P; P/S⇒P;
        对于 I=M 到 N 步长 1 执行
        A[I, J]-P*RE[I]⇒A[I, J]
      终;  注 {以上行变换}
      对于 I=1 到 N 步长 1 执行
      始 0⇒P;
        对于 J=N 到 M 步长 -1 执行
        P+RE[J]*A[I, J]⇒P; P/S⇒P;
        对于 J=M 到 N 步长 1 执行
        A[I, J]-P*RE[J]⇒A[I, J]
      终;  注 {以上列变换}
SKIP: Q⇒A[M, M-1]
    终;  注 {化为上海森堡型完毕}
    0⇒T; N⇒NN;
NW:  若 NN=0 则转 FIN 否;
    0⇒ITS; NN-1⇒N1;
  注 {以下检查是否有单个小的次对角线元素}
NT:  对于 L=NN 到 2 步长 -1 执行
      若 §ABS(A[L, L-1])≤EPS*(§ABS(A[L-1, L-1])+§ABS(A[L, L]))
      则始 L⇒LL; 转 CT1 终否;
      1⇒LL;
CT1:  A[NN, NN]⇒X;
      若 LL=NN 则转 ΘW 否;
      A[N1, N1]⇒Y; A[NN, N1]*A[N1, NN]⇒W;
      若 LL=N1 则转 TW 否;
      若 30≤ITS 则始停 888; 转 FIN 终否;
      若 ITS=10 则否若 ITS=20 则否转 SΘN;
      注 {以下作特殊移位}
      T+X⇒T;
      对于 I=1 到 NN 步长 1 执行 A[I, I]-X⇒A[I, I];
      §ABS(A[NN, N1])+§ABS(A[N1, NN-2])⇒S; 0.75*S⇒X; X⇒Y;
      -0.4375*S*S⇒W;
SΘN:  ITS+1⇒ITS;
      注 {以下检查是否有相邻两个小的次对角线元素}
      对于 M=NN-2 到 LL 步长 -1 执行
      始 A[M, M]⇒Z; X-Z⇒R; Y-Z⇒S;
      (R*S-W)/A[M+1, M]+A[M, M+1]⇒P; A[M+1, M+1]-Z-R-S⇒Q;
      A[M+2, M+1]⇒R; §ABS(P)+§ABS(Q)+§ABS(R)⇒S;
      P/S⇒P; Q/S⇒Q; R/S⇒R; M⇒MM;
```

```
        若 M=LL 则转 CT2 否;
        若 §ABS(A[M, M-1])*(§ABS(Q)+§ABS(R))⩽EPS*§ABS(P)*
          (§ABS(A[M-1, M-1])+§ABS(Z)+§ABS(A[M+1, M+1]))则转 CT2 否
      终;
CT2:  对于 I=MM+2 到 NN 步长 1 执行 0⇒A[I, I-2];
        对于 I=MM+3 到 NN 步长 1 执行 0⇒A[I, I-3];
        注 {以下对 LL 至 NN 行及 MM 至 NN 列进行双步 QR 变换}
        对于 K=MM 到 N1 步长 1 执行
        始若 K=MM 则否
          始 A[K, K-1]⇒P; A[K+1, K-1]⇒Q;
            若 K=N1 则 0⇒R 否 A[K+2, K-1]⇒R;
            §ABS(P)+§ABS(Q)+§ABS(R)⇒X;
            若 X=0 则转 CT3 否;
            P/X⇒P; Q/X⇒Q; R/X⇒R
          终;
          §SQRT(P*P+Q*Q+R*R)⇒S;
          若 P<0 则 -S⇒S 否;
          若 K=MM 则若 LL=MM 则否 -A[K, K-1]⇒A[K, K-1]
                    否 -S*X⇒A[K, K-1];
          P+S⇒P; P/S⇒X; Q/S⇒Y; R/S⇒Z; Q/P⇒Q; R/P⇒R;
          注 {以下行变换}
          对于 J=K 到 NN 步长 1 执行
          始 A[K, J]+Q*A[K+1, J]⇒P;
            若 K=N1 则否
            始 P+R*A[K+2, J]⇒P; A[K+2, J]-P*Z⇒A[K+2, J]
            终;
            A[K+1, J]-P*Y⇒A[K+1, J]; A[K, J]-P*X⇒A[K, J]
          终;
          若 K+3⩽NN 则 K+3⇒JJ 否 NN⇒JJ
          注 {以下列变换}
          对于 I=LL 到 JJ 步长 1 执行
          始 X*A[I, K]+Y*A[I, K+1]⇒P;
            若 K=N1 则否
            始 P+Z*A[I, K+2]⇒P; A[I, K+2]-P*R⇒A[I, K+2]
            终;
            A[I, K+1]-P*Q⇒A[I, K+1]; A[I, K]-P⇒A[I, K]
          终;
      CT3:
        终;
```

```
转 NT;
    注 {以下找到一个根的处理}
ΘW: X+T⇒RE[NN]; 0⇒IM[NN]; ITS⇒TS[NN]; N1⇒NN; 转 NW;
    注 {以下找到两个根的处理}
TW: (Y-X)/2⇒P; P*P+W⇒Q; §SQRT(§ABS(Q))⇒Y;
    -ITS⇒TS[NN]; ITS⇒TS[N1]; T+X⇒X;
    若 0<Q 则始若 P<0 则 -Y⇒Y 否;
                P+Y⇒Y; X+Y⇒RE[N1]; X-W/Y⇒RE[NN];
                0⇒IM[NN]; 0⇒IM[N1]
              终
            否始 X+P⇒RE[N1]; X+P⇒RE[NN];
                Y⇒IM[N1]; -Y⇒IM[NN]
              终;
    NN-2⇒NN; 转 NW;
FIN:
终;
```

参 考 资 料

[1] 北京大学等编,《计算方法》,人民教育出版社, 1961, 北京。

[2] Ф. Р. 甘特马赫著,柯召译,《矩阵论》,高等教育出版社, 1955。

[3] В. В. Воеводин, "Численные Методы Алгебры", Москва, 1966.

[4] Д. К. Фаддеев, В. Н. Фаддеева 著, 刘光武等译,《线性代数计算方法》,上海科技出版社, 1965.

[5] A. S. Householder, "The Theory of Matrices in Numerical Analysio" Blaisdell, New York, 1964.

[6] A. Ralston, H. S. Wilf, "Mathematical Methods for Digital Computers" Vol. I, II. (1960, 1967) John Wiley & Sons Inc. 第一卷有中译本: 徐献瑜等译,《数字计算机上的数学方法》,上海科学技术出版社, 1963。

[7] Schwarz/Rutishauser/Stiefel, "Matrizen-Numerik" B. G. Teubner, Stuttgart, 1972.

[8] J. H. Wilkinson, "Rounding Errors in Algebraic Processes" Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[9] J. H. Wilkinson, "The Algebraic Eigenvalue Problem" Clarendom Press, Oxford, 1965.

[10] J. H. Wilkinson, C. Reinsch, "Handbook for automatic Computation" Volume II. Linear algebra. Berlin-Heidelbesg-New York 1971.

[11] B. Parlett, "Global Convergence of the Basic QR Algorithm On Hessenberg Matrices" «Math. of Comput.» (1968) Vol. 22, pp. 803~817.

[12] В. В. Воеводин, "Решение Полной Проблемы Собственных Значений Обобщенным Методом Вращений" «Вычислительные Методы и Программирование» 3, 1965, Стр. 89~105.

[13] P. J. Eberlein, "A Jaobi-like method for the automatic computation of eigenvalues and eigenvectors of a arbitrary matrix" J. Soc. Indust. Appl. Math. 10 (1962) pp. 74~88.

[14] 周天孝,《数值确定矩阵全部特征值的一个 Jacobi 型方法》, 1965。

[15] G. Peters and J. H. Wilkinson, "$Ax=\lambda Bx$ and the generalized eigenproblem" «SIAM J. On Num. Analy». 7 (1970) pp. 479~492.

[16] G. Fix and R. Heiberger, "An Algorithm for the ill-Conditioned generalized eigenvalue problem" «SIAM J. On Num. Analy.» 9 (1972) pp. 78~88.

[17] G. peter and J. H. Wilkinson, "Eigenvalues of $Ax=\lambda Bx$ with band Symmetric A and B." «Computer J.» Vol. 12, 1969, pp. 398~404.

[18] K. K. Gupta, "Solution of eigenvalues problems by Sturm Sequence Methods" «Inter. J. for Num. Meth. in Eng.» Vol. 4(1972)pp. 379~404.

[19] "Optimal gradient minimization Schme for finite element eigenproblem"«J. Sound and Vibr.»20, 1972, No. 3, pp. 383~392.

[20] Kyuichiro Washizu, "Variational Methods in Elasticity and Plasticity". Pergamon Press Ltd, 1968.

第十一章　常微分方程初值问题数值解法

生产斗争和科学实验中有大量的问题都表达为常微分方程的形式。特别是在描述系统的动态演变时，例如生灭的过程、物体的运动、电路的振动瞬变、化学反应过程等等都表达为以时间 t 为基本变量参数的常微分方程或方程组(当然基本变量也可以表达空间或其它的物理意义)。

对于这类方程，实践上常常需要解"初值问题"，即给定了解的初始条件，以推算过程的发展，这就是本章的主题。另外也还要解所谓边值问题，即给定两端点条件以定解. 在简单情况下可以参照第十三章偏微分方程边值问题的解法；至于复杂的情况由于难度较大在本书中不作介绍。

为了解算常微分方程，应该对于它所反映的过程的基本特征有所了解，下面将通过一些简单的例子来介绍一些典型的特征。

§11.1　一些典型过程的微分方程

11.1.1　生灭过程与稳定性

考虑一阶常系数常微分方程

$$\frac{du}{dt}=\alpha u \tag{11.1.1}$$

这是一般的生灭、增长、蜕变等物理过程的最简单而最典型的数学模型。系数 α 为蜕变概率，例如在中子的生灭过程中 u 表示一定半径的球形裂变物质的中子总数，系数 α 表示蜕变率。

当初值 $u(0)=u_0$ 时，方程(11.1.1)的解为

$$u(t)=u_0e^{\alpha t} \tag{11.1.2}$$

当 $\alpha>0$ 时，$u(t)$ 随 $t\to\infty$ 而指数状 $\to\infty$；当 $\alpha<0$ 时 $u(t)$ 随 $t\to\infty$ 而指数状 $\to 0$；当 $\alpha=0$ 时 $u(t)\equiv u_0$ 即保持常值(图 11.1)。由此可见系数 α 的正或负刻划了过程的增或衰、生或灭的内在特征。例如在上述裂变过程中，当 $\alpha>0$ 时中子无穷增殖，即体系处于超临界状态。当 $\alpha<0$ 时中子数递减，过程趋于熄灭，即体系处于亚临界状态。当 $\alpha=0$ 时中子数不变，即处于临界状态。

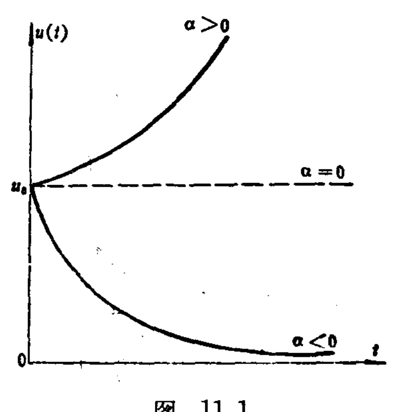


图 11.1

当 $\alpha<0$ 时命

$$\tau=\frac{1}{|\alpha|} \tag{11.1.3}$$

于是解(11.1.2)可以写成

$$u(t)=u_0e^{-t/\tau} \tag{11.1.4}$$

τ 具有时间的量纲叫做过程的"有生时间"，在物理和工

程的不同领域中也叫做“弛预时间”或“时间常数”等等。它刻划着蜕变的快慢，即表达了过渡历程的活跃时间。事实上，当 t 从 0 增至 τ 时 u 降至初值 u_0 的 $\frac{1}{e}=0.368$ 倍，而且当 t 进一步增至 2τ、3τ、$4\tau\cdots$ 则 u 更降至 u_0 的 $e^{-2}=0.135$、$e^{-3}=0.05$、$e^{-4}=0.018\cdots$ 倍。这就是说，当时间 t 超过 τ 的量级后过程即基本结束，而 τ 从量级上表达了过渡历程的活跃时间。因此“有生时间”τ 这样一个量有助于在实际问题中对于过程的发展作出基本的、概略的估计。

上述增或减、生或灭的特征也表现为过程对于小扰动是否为稳定。为了使讨论具有一定的意义，设(11.1.1)中的系数可以是复数，即考虑所谓“模型”方程

$$\frac{du}{dt}=\mu u,\quad \mu=\alpha+i\beta \tag{11.1.5}$$

设初值 u_0 受到小扰动，即 $u_0\sim u_0+\varepsilon_0$，则(11.1.5)的解 $u(t)$ 也相应受扰 $u(t)\sim u(t)+\varepsilon(t)$。由于受扰解同样也满足方程(11.1.5)即

$$\frac{d}{dt}(u+\varepsilon)=\mu(u+\varepsilon) \tag{11.1.6}$$

与(11.1.5)相减即得小扰动方程

$$\frac{d\varepsilon}{dt}=\mu\varepsilon \tag{11.1.7}$$

其解为

$$\varepsilon(t)=\varepsilon_0 e^{\mu t}=\varepsilon_0 e^{\alpha t}\cdot e^{i\beta t}=\varepsilon_0 e^{\alpha t}(\cos\beta t+i\sin\beta t) \tag{11.1.8}$$

注意，由于原始方程(11.1.5)是线性齐次的，小扰动方程(11.1.7)的形式和它一样。

当 $\mathrm{Re}\mu=\alpha<0$ 时，$\varepsilon(t)$ 随 $t\to\infty$ 而指数状衰减$\to 0$，即任何的初始微扰都会最终自动消失，这样的系统(11.1.5)是稳定的。反之，当 $\mathrm{Re}\mu=\alpha>0$，$\varepsilon(t)$ 随 $t\to\infty$ 而指数状$\to\infty$，即系统一旦受扰，就会越偏越远，这样的系统是不稳定的。当 $\mathrm{Re}\mu=\alpha=0$ 时，$\varepsilon(t)=\varepsilon_0 e^{i\beta t}$ 作简谐振动而幅度不变，这是临界稳定或中立稳定。由此可见，系统对于初始微扰的稳定性这一内在特征，取决于系数 μ 的实部为负或正。以后即将看到，复杂系统的判稳也正是基于这一简单的标准。

11.1.2 简谐振动和阻尼谐振

简谐振子的运动规律是

$$a\frac{d^2w}{dt^2}=-cw \tag{11.1.9}$$

w 为振子位移，$a>0$ 为质量，$c>0$ 为弹性常数，$-cw$ 为弹性恢复力，这里负号表示恢复力与位移反向。它的解为

$$w(t)=A\sin(\omega t+\varphi),\quad \omega=\sqrt{\frac{c}{a}} \tag{11.1.10}$$

常数 A(振幅)，φ(初始相角)可由两个初始条件 $w(0)$，$w'(0)$ 决定。ω 表示简谐振动的角频率，也可以用频率 $\nu=\frac{\omega}{2\pi}$ 或周期 $T=\frac{2\pi}{\omega}=\frac{1}{\nu}$ 来刻划这个振动特征。

解析上更方便些把解(11.1.10)表为复数形式

$$w(t)=a_1 e^{i\omega t}+a_2 e^{-i\omega t} \tag{11.1.11}$$

这里 $\pm i\omega$ 就是二阶微分方程(11.1.9)的二次本征方程的两个根:

$$\mu^2+\frac{c}{a}=0,\quad \mu_{1,2}=\pm i\omega=\pm i\sqrt{\frac{c}{a}} \tag{11.1.12}$$

在这里，由于本征根实部 $\mathrm{Re}\mu_{1,2}=0$，由(11.1.12)或(11.1.10)可以看出振动的幅度不变，因此是临界稳定的。

上面分举了增衰和谐振两种典型，实际这两者往往是耦合并存的。例如阻尼谐振方程为

$$a\frac{d^2w}{dt^2}+b\frac{dw}{dt}+cw=0 \tag{11.1.13}$$

这里增加了阻尼项 bw', $b\geqslant 0$, 在机械振动中这就是阻力，即正比而反向于速度。在电路振动中 $a=L$(电感)，$b=R$(电阻)，$c=\frac{1}{C}$，C(电容)。(11.1.13)的本征方程是

$$a\mu^2+b\mu+c=0 \tag{11.1.14}$$

$$\mu_{1,2}=-\frac{b}{2a}\pm\sqrt{\frac{b^2}{4a^2}-\frac{c}{a}}$$

$$=\begin{cases}-\dfrac{b}{2a}\pm i\dfrac{1}{2a}\sqrt{4ac-b^2}, & \text{为复数,当 } b^2<4ac\text{——低阻尼}\\[2ex] -\dfrac{b}{2a}\pm\dfrac{1}{2a}\sqrt{b^2-4ac}, & \text{为实数,当 } b^2\geqslant 4ac\text{——高阻尼}\end{cases} \tag{11.1.15}$$

而微分方程(11.1.13)的通解为

$$w(t)=a_1e^{\mu_1 t}+a_2e^{\mu_2 t} \tag{11.1.16}$$

由于本征根实部 $\mathrm{Re}\mu_{1,2}\leqslant 0$, 当低阻尼时，这是振幅衰退的振动过程，本征根虚部 $I_m\mu_{1,2}$ 给出其角频率，当高阻尼时则是纯衰减过程，都是稳定的。

在某些物理过程中，例如含有有源元件的线路中，可以有负电阻或负阻尼的作用，这时 $b<0$, 产生实部为正的本征根，即振幅随时间增大，这时就是不稳定的。

如命

$$w=u_1,\quad \frac{dw}{dt}=u_2$$

则一个二阶方程(11.1.13)等价于一阶方程组

$$\begin{aligned}\frac{du_1}{dt}&=a_{11}u_1+a_{12}u_2,\\ \frac{du_2}{dt}&=a_{21}u_1+a_{22}u_2,\end{aligned}\qquad \boldsymbol{A}=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}=\begin{bmatrix}0 & 1\\ -\dfrac{c}{a} & -\dfrac{b}{a}\end{bmatrix} \tag{11.1.17}$$

系数矩阵 $\boldsymbol{A}$ 的本征方程

$$|\boldsymbol{A}-\mu\boldsymbol{I}|=\begin{vmatrix}-\mu & 1\\ -\dfrac{c}{a} & -\dfrac{b}{a}-\mu\end{vmatrix}$$

$$=\mu^2+\frac{b}{a}\mu+\frac{c}{a}=0 \tag{11.1.18}$$

顺便再回到一阶模型微分方程(11.1.5)，那里的系数 μ 就是其相应的一次代数本征方程的根，因此系统的判稳实质是针对着本征根。在这里也可以理解，为什么在模型方程(11.1.5)考虑到复系数，因为这样在实际上就已经把谐振和阻尼谐振包括在内了。

§ 11.2 一般的微分方程组及其稳定性

11.2.1 常系数线性微分方程组

$$\frac{du_j}{dt}=\sum_{k=1}^{m}a_{jk}u_k+g_j(t),\quad j=1,\cdots,m \tag{11.2.1}$$

或者采用向量的写法

$$\boldsymbol{u}(t)=\boldsymbol{u}=\begin{bmatrix}u_1\\ \vdots\\ u_m\end{bmatrix},\quad \boldsymbol{g}(t)=\boldsymbol{g}=\begin{bmatrix}g_1\\ \vdots\\ g_m\end{bmatrix},\quad \boldsymbol{A}=\begin{bmatrix}a_{11}\cdots a_{1m}\\ \vdots\\ a_{m1}\cdots a_{mm}\end{bmatrix}$$

这就简化为

$$\frac{d\boldsymbol{u}}{dt}=\boldsymbol{A}\boldsymbol{u}+\boldsymbol{g} \tag{11.2.2}$$

系数矩阵 $\boldsymbol{A}$ 决定系数的内在特征，右端 $\boldsymbol{g}(t)$ 为外部驱动项。

分析系统稳定性的方法基本同于前。设初值为 $\boldsymbol{u}_0$ 时介向量为 $\boldsymbol{u}(t)$，而初值受扰 $\boldsymbol{u}_0\sim\boldsymbol{u}_0+\varepsilon_0$ 时受扰介为 $\boldsymbol{u}(t)+\varepsilon(t)$，即

$$\frac{d}{dt}(\boldsymbol{u}+\varepsilon)=\boldsymbol{A}(\boldsymbol{u}+\varepsilon)+\boldsymbol{g} \tag{11.2.3}$$

与(11.2.2)相减即得所谓的小扰动方程

$$\frac{d\varepsilon}{dt}=\boldsymbol{A}\varepsilon \tag{11.2.4}$$

它在形式上与原始方程(11.2.2)基本一样，只是没有驱动项，即方程组是齐次的。

设矩阵 $\boldsymbol{A}$ 具有互异的本征值 $\mu_1,\cdots,\mu_m$，则 $\boldsymbol{A}$ 可以化为对角型㊀，即有非异矩阵 $\boldsymbol{P}$ 使得

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}=\boldsymbol{M}=\begin{bmatrix}\mu_1 & & 0\\ & \ddots & \\ 0 & & \mu_m\end{bmatrix}$$

这里 $\boldsymbol{P}$ 的各个列向量就是对应于各个本征值的本征向量。于是作变换

$$\boldsymbol{\xi}=\boldsymbol{P}^{-1}\varepsilon\quad 即\quad \varepsilon=\boldsymbol{P}\boldsymbol{\xi}$$

则方程组(11.2.4)等价于对角型的方程组

$$\frac{d\boldsymbol{\xi}}{dt}=\boldsymbol{M}\boldsymbol{\xi}$$

亦即可以"分解"为互相独立的方程：

$$\frac{d\xi_j}{dt}=\mu_j\xi_j,\quad j=1,\cdots,m$$

这里每一个方程都是像模型方程(11.1.5)那样最简单的形式，系数 μ_j 可以是复数。

而扰动(用变换后的 ξ 表示)可以表为

$$\xi_j(t)=\xi_{j0}e^{\mu_j t}=\xi_{j0}e^{\alpha_j t}\cdot e^{i\beta_j t},\quad j=1,\cdots,m$$

此处

$$\alpha_j=\mathrm{Re}\mu_j,\quad \beta_j=I_m\mu_j$$

㊀ 这里为了简化起见，不讨论一般的情况，那时虽不一定能化为对角型，但总可以化为类似于对角型的所谓若当标准型的。

由此可见系数矩阵 $\boldsymbol{A}$ 的本征值的实部对应于振幅的增减，虚部对应于振动的角频率。因此当本征值的实部 $\mathrm{Re}\mu_1, \cdots, \mathrm{Re}\mu_m$ 均 <0 时系统是稳定的，即任何初始扰动都随 t 的增长而衰退。这时命

$$\tau_j=\frac{1}{|\mathrm{Re}\mu_j|}, \quad j=1, \cdots, m \tag{11.2.5}$$

也称为有生时间或时间常数。在这里系统是由若干“成分”组合而成，不同的“成分”可以有不同的时间常数。其中最大的时间常数 $\tau_{\max}=\max\tau_j$ 表达了全过程的活跃时间；而最小的时间常数 $\tau_{\min}=\min\limits_j \tau_j$ 则表达系统最“敏感”环节的反应速度，即过渡时间。在许多领域里常常出现最大的时间常数与最小的时间常数相差悬殊，例如达到若干个数量级的情况，这样的系统或微分方程组称为病态的(stiff)。

如果矩阵 $\boldsymbol{A}$ 的某些本征值的实部 $\mathrm{Re}\,\mu_j=0$，则这些“成分”相应于以虚部 $I_m\mu_j=\beta_j$ 为角频率的无阻尼简谐振动，如 11.1.2 节所述。如有某些 $\mathrm{Re}\mu_j>0$，则一旦受扰后会随 t 的增长而无穷增长，即过程是不稳定的。

11.2.2 变系数及非线性微分方程组

在变系数线性系统

$$\frac{d\boldsymbol{u}}{dt}=\boldsymbol{A}(t)\boldsymbol{u}(t)+\boldsymbol{g}(t) \tag{11.2.6}$$

系数矩阵 $\boldsymbol{A}=\boldsymbol{A}(t)$ 依赖于 t，它的本征值也就依赖于 t，$\mu_j=\mu_j(t)$。小扰动方程则为

$$\frac{d\boldsymbol{\varepsilon}}{dt}=\boldsymbol{A}(t)\boldsymbol{\varepsilon}$$

因此系统的稳定性依赖于所处的时段，情况就复杂些。

在非线性系统，一般表为

$$\frac{du_j}{dt}=f_j(u_1, \cdots, u_m, t), \quad j=1, \cdots, m \tag{11.2.7}$$

右端 f_j 可以是 $u_1, \cdots, u_m$ 以及 t 的非线性函数，它可以简写为向量的形式

$$\frac{d\boldsymbol{u}}{dt}=\boldsymbol{f}(\boldsymbol{u}, t), \quad \boldsymbol{u}=\begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \quad \boldsymbol{f}=\begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix}$$

可以采用所谓线性化的方法来分析它对于扰动的稳定性。设 $\boldsymbol{u}$(即 $u_1, \cdots, u_m$) 是一组特解，$\boldsymbol{u}+\boldsymbol{\varepsilon}$(即 $u_1+\varepsilon_1, \cdots, u_m+\varepsilon_m$) 是相应的受扰解。

由于

$$\frac{d}{dt}(u_j+\varepsilon_j)=f_j(u_1+\varepsilon_1, \cdots, u_m+\varepsilon_m, t)$$

视 $\varepsilon_1, \cdots, \varepsilon_m$ 为微量，作线性幂次展开

$$f_j(u_1+\varepsilon_1, \cdots, u_m+\varepsilon_m, t)\approx f_j(u_1, \cdots, u_m, t)+\sum_{k=1}^{m}\frac{\partial f_j}{\partial u_k}\varepsilon_k \tag{11.2.8}$$

因此扰动 $\varepsilon_1, \cdots, \varepsilon_m$ 近似地满足齐次线性微分方程组

$$\frac{d\boldsymbol{\varepsilon}}{dt}=\boldsymbol{A}\boldsymbol{\varepsilon}, \quad \boldsymbol{A}=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}=\begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_m} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial u_1} & \cdots & \frac{\partial f_m}{\partial u_m} \end{bmatrix} \tag{11.2.9}$$

这里 $A=\frac{\partial f}{\partial u}$ 为导数矩阵(亦称雅可比 Jacobi 阵，注意它是在特定解 $u_1, \cdots, u_m$ 处计算)。方程组(11.2.9)称为原来的非线性方程组(11.2.7)的线性化小扰动方程。它的判稳仍按照前述本征值分析方法。由于本征值 μ 不仅依赖于所处的时刻，而且依赖于特解 $u_1, \cdots, u_m$，情况也就更复杂了。当方程为线性时导数矩阵就是系数矩阵。

像(11.2.7)这样的一阶方程组的形式已经足够一般，以至于无须专门去论列高阶微分方程或方程组的情况。事实上，除了一些特殊情况外，仿照 §11.1 末列举的方法，总可以把高阶方程化成一阶微分方程组(11.2.7)。这样做对解析处理和数值处理都有统一简化的好处，在计算实践中通常也就是这样办的。

11.2.3 病态微分方程

上面对线性方程提到了病态的问题，这个概念自然也适用于一般非线性方程，即以导数矩阵的最大最小时间常数的比值 $\tau_{\max}/\tau_{\min}$ 作为病态程度的衡量。在实践中常常出现同一系统内有相差悬殊的时间常数。例如在自动控制系统中控制线路常常是反应灵敏的，能够迅速完成状态过渡，即具有小的时间常数；反之，受控体本身的运动由于惯性较大，状态过渡较慢，即具有大的时间常数。在多组分化学反应中，可能有些反应速度很快，有些很慢，在电力网络、电子网络或者由扩散、传热过程对空间变量离散化而得常微分方程中也经常出现量级悬殊的时间常数。

最大时间常数表达了全过程的活跃时间，即决定积分求解的总时间；而在经典的数值解法中积分步长则受限于最小时间常数(见 §11.4)。于是当病态度很大时，积分步数往往很大，这是病态方程带来的计算困难。

在实际系统中往往有这样的元件或环节，它对全过程并无实质性的影响，但却带来极小的时间常数，导致病态化，造成计算的累赘。因此在形成数学问题时应作物理和数学的分析，对有关因素适当地权衡取舍，以避免不必要的病态性。

§11.3 差分方法和有关的概念

在各项科学技术领域中，常微分方程问题经常出现，有大量的求解需要，同时，计算机和数值方法的发展也提供了解决这类问题的实际可能，因此常微分初值问题在计算实践中占有很大的比重。

我们知道，传统的数学分析方法只能解决少数的比较简单和典型的常微分方程问题，比如说一般只能胜任常系数线性方程，对于变系数线性方程就有很大困难，更不用说一般的非线性方程了。一般说来解析方法比较适合于定性的研究，当然也还远不能解决问题。数值方法的实用范围则远为宽广。对于绝大多数实践上出现的常微分方程初值问题，无论是常系数还是变系数，是线性还是非线性，一般都能应用数值方法在实际上得到解决。在生产实践上，对于非线性和复杂系统的问题，应用计算机和数值方法的效果最为显著，因为正是在这个方面传统的数学分析方法是难以胜任的。

常微分方程初值问题数值解的主要手段是差分方法，它的优点是通用性强即适应面广而方法简单便于掌握。在这一章中，先通过最简单的一类差分方法，即尤拉方法来说明一些有关的基本概念，包括数值稳定性等问题，见§11.3～5，解法的进一步的介绍则见§11.6～9。

11.3.1 尤拉方法

设有一阶常微分方程组的初值问题

$$\begin{cases} \boldsymbol{u}'=\boldsymbol{f}(\boldsymbol{u},\ t) \\ \boldsymbol{u}(0)=\boldsymbol{u}_0 \end{cases} \tag{11.3.1}$$

在这里和以后，为了简便主要将采用向量写法。

将自变量 t 离散化，取步长 h，命

$$t_n=nh,\ \boldsymbol{u}_n=\boldsymbol{u}(t_n),\ \boldsymbol{f}_n=\boldsymbol{f}(\boldsymbol{u}_n,\ t_n),\quad n=0,\ 1,\ \cdots$$

用适当的差商代替(11.3.1)的左端 $\boldsymbol{u}'$，用适当的 f 值的组合代替右端的 $\boldsymbol{f}(\boldsymbol{u},\ t)$，很自然地可以得到一系列的差分格式，称为尤拉公式如下：

(1) $\frac{1}{h}(u_{n+1}-u_n)=f_n$, 向前差公式 (11.3.2)

(2) $\frac{1}{h}(u_{n+1}-u_n)=f_{n+1}$, 向后差公式 (11.3.3)

(3) $\frac{1}{h}(u_{n+1}-u_n)=\frac{1}{2}(f_{n+1}+f_n)$, 平均(梯形)公式 (11.3.4)

(4) $\frac{1}{2h}(u_{n+1}-u_{n-1})=f_n$ 中心差公式 (11.3.5)

公式(11.3.2)中导数在点 t_n 计算，而差商取 u_n 及向前一点 u_{n+1}，因此叫做向前差公式；在公式(11.3.3)中导数则在点 t_{n+1} 计算，而差商取 u_{n+1} 及向后一点 u_n，因此叫向后差公式。

我们将通过它们来说明有关差分方法的几个概念，如截断误差，显式和隐式，单步法和多步法以及数值稳定性。

11.3.2 截断误差

将向前差分公式(11.3.1)的左右两端同在 $t=t_n$ 作幂次展开，由于

$$\begin{aligned}\frac{1}{h}(u_{n+1}-u_n)&=u_n'+\frac{h}{2}u_n''+O(h^2)\\&=u_n'+O(h)\end{aligned}$$

因此得差分方程与微分方程在 t_n 点的“差”

$$E=\left\{\frac{1}{h}(u_{n+1}-u_n)-f_n\right\}-\{u_n'-f_n\}=O(h)$$

这个量 E 叫做截断误差。于是，当 $h\to0$ 时 $E\to0$，即差分方程(11.3.2)的极限形式就是微分方程(11.3.1)。因此可以认为差分方程(11.3.2)是微分方程(11.3.1)的一个合理的逼近。类似地对于向后差分公式(11.3.3)在 $t=t_{n+1}$ 作幂次展开，则得截差 $E=O(h)$。

对于公式(11.3.4)、(11.3.5)分别在

$$t=t_{n+\frac{1}{2}}=\frac{1}{2}(t_n+t_{n+1})$$

以及 $t=t_n$ 展开则都得截差 $E=O(h^2)$，因此它们相对于(11.3.2)和(11.3.3)而言精度提高一阶。

当差分公式的截差 $E=O(h^p)$ 时，我们说它具有 p 阶精度。

11.3.3 显式和隐式

解初值问题的差分方法的共同算法特点是步进式，即从初始一点或几点出发，每一步根据 u_n 或(及)其前的 u_{n-1}, …来计算新的 u_{n+1}，这样逐步推进。但是在每步的算法执行上还有差别。

例如已知了 u_n，因此也有了 $f_n=f(u_n, t_n)$，则根据公式(11.3.2)立即可以明显得出 u_{n+1}

$$u_{n+1}=u_n+hf_n$$

类似地，已知了 u_n, u_{n-1}，因此也有了 f_n, f_{n-1}，则根据公式(11.3.5)立即明显地得 u_{n+1}

$$u_{n+1}=u_{n-1}+2hf_n$$

这样的格式称为显式。

反之，在公式(11.3.3)和(11.3.4)中除了显含 u_{n+1} 以外，在 $f_{n+1}=f(u_{n+1}, t_{n+1})$ 中还隐含未知的 u_{n+1}，因此必须"解"方程才能得出 u_{n+1}，这样的格式称为隐式。

从算法上说，显式远比隐式方便。但有时由于其它因素，如精度，特别是稳定性的考虑，有时宁愿采用隐式(见后 §11.4 之末)。

11.3.4 单步与多步

步进差分格式除了显、隐之别外还有所谓单步和多步的差别。例如在公式(11.3.2)、(11.3.3)及(11.3.4)，当从 t_n 推进到 t_{n+1} 时只用到当前时刻 t_n 的数据，因此称为单步的；反之，在公式(11.3.5)中则要时刻 t_n 及 t_{n-1} 的数据，因此称为双步的，这是多步的一种，可以有三步及更多步的。

从存储量来看，多步法需要保留多个时刻的数据，因此是不利的，但当方程个数即未知函数个数不太巨大时，问题并不严重。

单步法还有可以自动起步及自由改变步长的优点；多步法则不便。例如在公式(11.3.5)除了 u_0 外还需要 u_1 才能起动，通常可用单步法起步或采取其它算法措施以补足必须的初始几点值。

线性差分格式(包括显、隐、单步及多步)的一般形式是

$$\frac{1}{h}(\alpha_0u_{n+1}+\alpha_1u_n+\cdots+\alpha_ku_{n+1-k})=\beta_0f_{n+1}+\beta_1f_n+\cdots+\beta_kf_{n+1-k} \tag{11.3.6}$$

这里 α_0, …, α_k, β_0, …, β_k 都是不依赖于具体微分方程的常数。可以看到(11.3.2)～(11.3.5)都是其特例。当 $\beta_0=0$ 时是显式，否则是隐式。当 $k\geqslant2$ 而 α_k 和 β_k 不同为0时是多步的，当 $k<2$ 时是单步的。

这里称(11.3.6)为线性格式是指它表示为 u_{n+1}, …, u_{n+1-k}, f_{n+1}, …, f_{n+1-k} 的线性组合的形式。也有非线性的差分格式见 §11.8。

§11.4 数值稳定性

上面列举的差分格式(11.3.2)～(11.3.5)都是微分方程组(11.3.1)的合理逼近，当 $h\to0$ 时截差 $E\to0$，有的还达到较高精度，如(11.3.4)、(11.3.5)的截差 $E=O(h^2)$。但是，单靠这一点还不足以保证差分格式可以工作。关键的问题是差分方程是否具有对于扰动的

稳定性，即数值稳定性。在数值不稳定的情况下，计算误差将恶性发展以致计算失败。

11.4.1 判稳方法

我们试用前列格式解模型微分方程

$$u'=\mu u, \quad \mu=\alpha+i\beta \tag{11.4.1}$$

并设 $\operatorname{Re}\mu=\alpha<0$，即微分方程是稳定的。这就相当于将一般的微分方程 $u'=f(u, t)$ 的右项 f 取为 μu，因此差分方程(11.3.6)成为线性齐次差分方程

$$\frac{1}{h}(\alpha_0 u_{n+1}+\alpha_1 u_n+\cdots+\alpha_k u_{n+1-k})=\mu(\beta_0 u_{n+1}+\beta_1 u_n+\cdots+\beta_k u_{n+1-k})$$

设 $u_j(j=0, 1, 2, \cdots)$ 是一个解。另设 $u_j+\varepsilon_j$ 是一个受扰解，即

$$\frac{1}{h}(\alpha_0(u_{n+1}+\varepsilon_{n+1})+\cdots+\alpha_k(u_{n+1-k}+\varepsilon_{n+1-k}))$$

$$=\mu(\beta_0(u_{n+1}+\varepsilon_{n+1})+\cdots+\beta_k(u_{n+1-k}+\varepsilon_{n+1-k}))$$

两式相减即得小扰动差分方程

$$\alpha_0\varepsilon_{n+1}+\alpha_1\varepsilon_n+\cdots+\alpha_k\varepsilon_{n+1-k}=\mu h(\beta_0\varepsilon_{n+1}+\beta_1\varepsilon_n+\cdots+\beta_k\varepsilon_{n+1-k}) \tag{11.4.2}$$

亦即

$$(\alpha_0-\mu h\beta_0)\varepsilon_{n+1}+(\alpha_1-\mu h\beta_1)\varepsilon_n+\cdots+(\alpha_k-\mu h\beta_k)\varepsilon_{n+1-k}=0$$

这是 $k+1$ 项的线性齐次差分方程，它的通解是㊀

$$\varepsilon_n=a_1\lambda_1^n+a_2\lambda_2^n+\cdots+a_k\lambda_k^n$$

此处 $\lambda_1, \cdots, \lambda_k$(均依赖于 μh)是对应于(11.4.2)的 k 次本征方程

$$\alpha_0\lambda^k+\alpha_1\lambda^{k-1}+\cdots+\alpha_k=\mu h(\beta_0\lambda^k+\beta_1\lambda^{k-1}+\cdots+\beta_k) \tag{11.4.3}$$

亦即

$$(\alpha_0-\mu h\beta_0)\lambda^k+(\alpha_1-\mu h\beta_1)\lambda^{k-1}+\cdots+(\alpha_k-\mu h\beta_k)\lambda^0=0$$

的 k 个根。显然，数值稳定，即扰动不随 n 增长的条件是

$$|\lambda_j(\mu h)|\leqslant 1, \quad j=1, \cdots, k \tag{11.4.4}$$

否则扰动 ε_n 将随 $n\to\infty$ 而无穷增长，这就是数值不稳定。

可以采用几何的方法来判稳。事实上，方程(11.4.3)表达微分方程本征值与步长的乘积 μh 与差分方程本征值 λ 之间的关系。为了方便，命复变量

$$z=\mu h, \quad z=x+iy, \quad \mu=\alpha+i\beta$$

于是本征方程(11.4.3)可以写成

$$\mu h\equiv z=\frac{\alpha_0\lambda^k+\alpha_1\lambda^{k-1}+\quad+\alpha_k}{\beta_0\lambda^k+\beta_1\lambda^{k-1}+\cdots+\beta_k}\equiv f(\lambda) \tag{11.4.5}$$

它表示了复数 λ 平面与复数 $z=\mu h$ 平面之间的变换。z 平面内所有满足条件(11.4.4)的点 $z=\mu h$ 的集合 Ω 叫做差分方程的稳定域。当微分方程的所有本征值 μ 乘以步长 h 后都落在 Ω 内则数值稳定。为了搞清楚稳定域 Ω 可以先研究临界即 $|\lambda|=1$ 的情况。变换(11.4.5)把 λ 平面上的单位圆 $|\lambda|=1$(即 $\lambda=e^{i\theta}$)，映为 z 平面上的一条临界曲线

$$\Gamma: z=f(e^{i\theta}) \quad 0\leqslant\theta\leqslant 2\pi$$

这是一条封闭的或展至无穷的并且可以有自相交的曲线，它把 z 平面分割为有限多个子域 $\Omega_1, \cdots, \Omega_r$。然后检查其中哪一些能确保(11.4.4)。

㊀ 为了简便，不去涉及有重根的一般情况。

11.4.2 尤拉公式的稳定性

(1) 向前差公式：$z=\lambda-1$

临界曲线 Γ：$|\lambda|=1$ 应为中心在 $z=-1$、半径为 1 的圆，其内部 Ω 是稳定域 $|\lambda|\leqslant 1$，如图 11.2。显然，对于任何 $\mathrm{Re}\,\mu<0$（即微分方程为稳定），只须取 h 充分小必能使 μh 落在稳定域之内。稳定的解析条件显然是

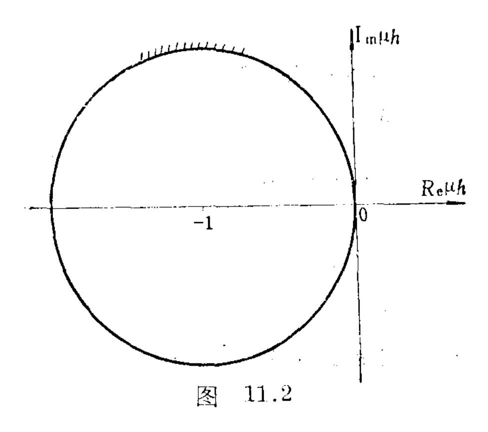


图 11.2

$$
\begin{aligned}
1\geqslant|\lambda|^2=|1+z|^2&=|1+\mu h|^2\\
&=(1+\alpha h)^2+(\beta h)^2\\
&=1+2\alpha h+(\alpha^2+\beta^2)h^2\\
&=1-2|\alpha|h+(\alpha^2+\beta^2)h^2
\end{aligned}
$$

即

$$h\leqslant\frac{2|\alpha|}{\alpha^2+\beta^2}=\frac{2|\mathrm{Re}\,\mu|}{|\mu|^2}\tag{11.4.6}$$

如果 $|\mathrm{Re}\mu|\gg|I_m\mu|$，则 $\dfrac{|\mathrm{Re}\,\mu|}{|\mu|^2}\approx\dfrac{1}{|\mathrm{Re}\,\mu|}=\tau$ 即时间常数，则稳定条件近似地表为

$$h\leqslant 2\tau\tag{11.4.7}$$

由此可见，即使原微分方程为稳定，即 $\mathrm{Re}\,\mu<0$，但 h 相当大以致于 μh 落在 Ω 之外，差分格式是不稳定的。反之，对于任何 $\mathrm{Re}\,\mu<0$，只要取 h 充分小，总能使 μh 落在 Ω 之内，这时差分格式为稳定。因此我们说这样的差分格式（相对于稳定的微分方程而言）为条件稳定。

(2) 向后差公式：$z=1-\lambda^{-1}$

临界曲线 Γ 是中心在 $z=1$ 半径为 1 的圆，其外 Ω 为稳定域（图 11.3），稳定域包含了全部左半平面。因此对于稳定的微分方程（即 $\mathrm{Re}\,\mu<0$），不论取步长 h 如何差分格式总是稳定的。这样的格式称为（对于稳定的微分方程而言）恒稳或无条件稳定的。

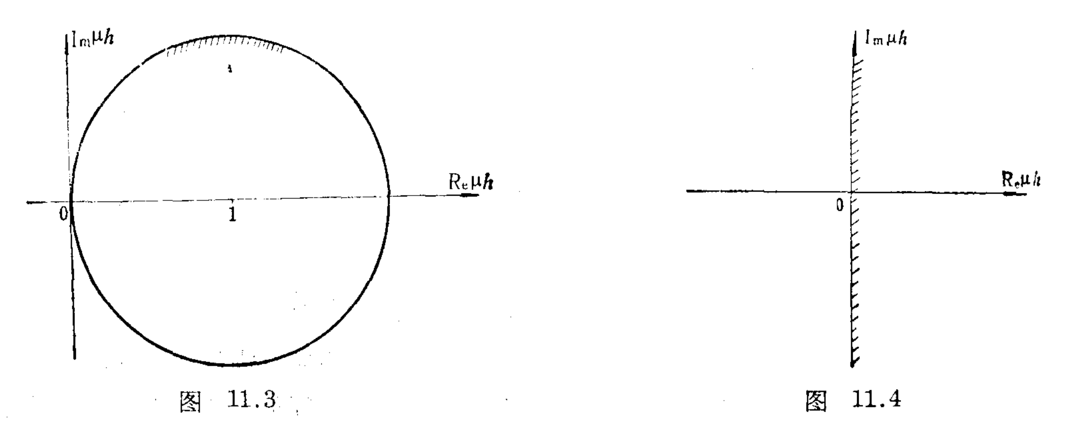


图 11.3　　图 11.4

(3) 平均公式：$z=2\dfrac{\lambda-1}{\lambda+1}$

临界曲线 Γ 为虚轴，其左半平面为稳定域 Ω（图 11.4）。这也是恒稳的。注意，当 $\mathrm{Re}\,\mu=0$，即微分方程为临界稳时，$|\lambda_{1,2}|=1$ 即差分方程也是临界稳的。

(4) 中心差公式：$z=\dfrac{1}{2}\left(\lambda-\dfrac{1}{\lambda}\right)$

临界曲线 Γ 为虚轴上的一个"切口"$[-i,\ i]$，稳定域 Ω 在其内部，实际上不存在（图

11.5)。因此,对任何 $\mathrm{Re}\,\mu<0$, 不论步长 h 如何,差分格式不稳定,可以称为(对于稳定的微分方程而言)恒不稳格式。因此一般说来是不实用的。但是对于临界稳的微分方程,即 $\mathrm{Re}\,\mu=0$,则差分方程 $|\lambda_{1,2}|=1$,即保持临界稳定性,故对于如无阻尼的简谐运动方程则是合适的。

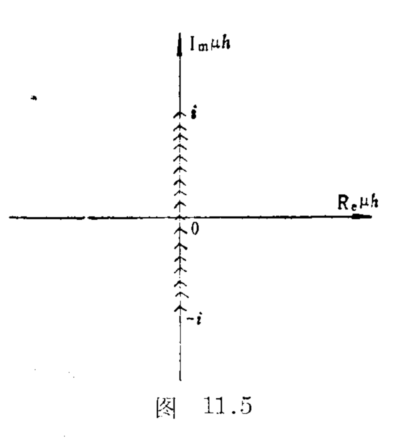


图 11.5

上面的例子说明了微分方程的稳定性与相应的差分方程的稳定性是互相联系而又有区别的。对于稳定的微分方程(大多数实际求解的问题都属此类)可以有条件稳、恒稳乃至恒不稳的差分格式。由于微分方程的稳定性是系统的一个重要的内在特征,因此在差分化时应该要求保持这个特征,即要求差分方程为稳定。事实上不稳定的差分方程(包括由步长过大引起的情况)是不能据以工作的。在计算中数值不稳定往往会很快很尖锐地表现出来,例如数值单调地或波动地迅速增大导致上溢等等,这时应该根据求解的微分方程和差分方法的特性而调整步长或更换格式。

11.4.3 非线性方程差分解法的判稳问题

上面是通过解一个模型方程 $u'=\mu u$ 来对差分格式进行判稳,初看起来似有很大局限性,其实,不完全是这个样子。设用多步格式(11.3.6)来解一般的非线性方程组

$$\boldsymbol{u}'=f(\boldsymbol{u},\ t)$$

即得非线性的差分方程组

$$\alpha_0 u_{n+1}+\cdots+\alpha_k u_{n+1-k}=h(\beta_0 f(u_{n+1},\ t_{n+1})+\cdots+\beta_k f(u_{n+1-k},\ t_{n+1-k}))$$

设 $u_j\,(j=0,\ 1,\ 2,\ \cdots)$ 为一组特解, $u_j+\varepsilon_j\,(j=0,\ 1,\ \cdots)$ 为一组相应的受扰解,则用类似于 11.2.2 节末的方法可知 ε_j 满足线性化小扰动差分方程组

$$\alpha_0\boldsymbol{\varepsilon}_{n+1}+\cdots+\alpha_k\boldsymbol{\varepsilon}_{n+1-k}=h(\beta_0\boldsymbol{A}_{n+1}\boldsymbol{\varepsilon}_{n+1}+\cdots+\beta_k\boldsymbol{A}_{n+1-k}\boldsymbol{\varepsilon}_{n+1-k}) \tag{11.4.8}$$

此处

$$\boldsymbol{A}_{n+1}=\left(\frac{\partial\boldsymbol{f}(\boldsymbol{u},\ t)}{\partial\boldsymbol{u}}\right)_{\boldsymbol{u}=\boldsymbol{u}_n,\ t=t_n}$$

即导数矩阵,计算于点 $\boldsymbol{u}=\boldsymbol{u}_n$, $t=t_n$。为了判稳可以把 $\boldsymbol{A}_{n+1}$, $\boldsymbol{A}_n$, …“冻结”,即认为 $\boldsymbol{A}_{n+1}\approx\boldsymbol{A}_n\approx\cdots\approx\boldsymbol{A}_{n+1-k}\approx\boldsymbol{A}$(事实上许多问题中导数矩阵也确是缓变的),于是(11.4.8)简化为

$$\alpha_0\boldsymbol{\varepsilon}_{n+1}+\cdots+\alpha_k\boldsymbol{\varepsilon}_{n+1-k}=h\boldsymbol{A}(\beta_0\boldsymbol{\varepsilon}_{n+1}+\cdots+\beta_k\boldsymbol{\varepsilon}_{n+1-k}) \tag{11.4.9}$$

同样用 11.2.1 节中的方法,把矩阵 $\boldsymbol{A}$ 对角化

$$\boldsymbol{A}\sim\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}=\boldsymbol{M}=\begin{bmatrix}\mu_1 & & 0\\ & \ddots & \\ 0 & & \mu_m\end{bmatrix}$$

并对 $\boldsymbol{\varepsilon}$ 作变换 $\boldsymbol{P}^{-1}\boldsymbol{\varepsilon}$,则方程组分解为 m 个独立的方程,其中每一个均如(11.4.1)那样模型方程的形式,而 μ 分别为 $\boldsymbol{A}$ 的本征值 μ_1, …, μ_m。这样,差分格式的判稳条件就成为要求 $\mu h\,(\mu=\mu_1,\ \cdots,\ \mu_m)$ 均落在稳定域之内。

对于稳定的方程组有

$$\mathrm{Re}\,\mu_j<0,\quad j=1,\ 2,\ \cdots,\ m$$

于是尤拉向前差格式的稳定条件(见(11.4.5)、(11.4.7))

$$h\leqslant\frac{2|\mathrm{Re}\,\mu_j|}{|\mu_j|^2},\quad j=1,\ 2,\ \cdots,\ m$$

亦即

$$h\leqslant\min_j\frac{2|\mathrm{Re}\,\mu_j|}{|\mu_j|^2}$$

或近似地,当$|\mathrm{Re}\mu_j|\gg|I_m\mu_j|$时,

$$\frac{|\mathrm{Re}\,\mu_j|}{|\mu_j|^2}\approx\frac{1}{|\mathrm{Re}\,\mu_j|}=\tau_j$$

$$h\leqslant\min_j 2\tau_j=2\tau_{\min}\tag{11.4.10}$$

即步长要取相当于最小时间常数的量级。其他一些经典显式方法如亚丹斯法、龙格-库塔方法的稳定条件也基本上是这个量级(见 11.6.1 节, 11.6.3 节)。条件(11.4.10)往往是一个极其苛刻的条件,即单纯从截断误差来考虑步长往往无须这样小,而从稳定性的要求则必须服从这个条件。

至于恒稳的格式,在前面是相对于模型方程而言的,即对一切 $\mathrm{Re}\,\mu<0$ 及一切 h 均稳定,那末对于有限多个 $\mathrm{Re}\,\mu_j<0$ 而言恒稳的结论自然不变。

在 §11.2 中提到,在实践中常常出现所谓病态微分方程,即最大与最小时间常数之比 $\tau_{\max}/\tau_{\min}$ 很大,达到 10^4, 10^5 乃至更高的量级。我们知道(§11.2)最大的时间常数 $\tau_{\max}$ 表达了全过程的活跃时间,因此积分求解的时段只须取为 $\tau_{\max}$ 或其数倍,相当或稍大半个量级。反之,如上所述,对尤拉向前差分或其它一些主要的经典显式方法而言,积分步长为最小时间常数的量级。因此积分的总步数将达 $\tau_{\max}/\tau_{\min}$ 或略高的量级。因此即使当方程组本身很简单,但病态度很大时,计算工作量可以很沉重甚至成为难以容忍的负担。困难的根源是显式步长条件,这是由格式的稳定性引起的。

相反地,恒稳格式如尤拉向后差及平均公式等等,由于它们的步长只受截断误差的制约而摆脱了稳定性制约,因此对于解病态方程是特别有利的。用向后差隐式比用向前差显式有时步长可以放大到几个量级。但是隐式方法每一步有一个解方程的问题,需要认真对待。

§11.5 隐式方程和相应解法

比较尤拉向前、向后及平均格式的稳定性及截断误差,可知隐式可以起到改善稳定性和(或)提高精度的作用。对于病态方程而言,改善稳定性则是主要的。对于非病态问题则提高精度是主要的。但是,隐式的每一步要解一个以 $\boldsymbol{u}_{n+1}$ 为未知向量的非线性或线性代数方程组。其解法的选择关系很大。

我们以最简单向后差隐式为例,介绍两种基本的方法。这时每步要解方程组(m 个方程)

$$\boldsymbol{u}_{n+1}-\boldsymbol{u}_n=h\boldsymbol{f}(\boldsymbol{u}_{n+1},\ t_{n+1})\tag{11.5.1}$$

未知向量为 $\boldsymbol{u}_{n+1}$。

11.5.1 比卡迭代法和预估校正公式

所谓比卡(Picard)迭代法就是

$$\boldsymbol{u}_{n+1}^{(j+1)}-\boldsymbol{u}_n=h\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(j)},t_{n+1})\quad j=1,2,\cdots\tag{11.5.2}$$

迭代直到

$$|\boldsymbol{u}_{n+1}^{(j+1)}-\boldsymbol{u}_{n+1}^{(j)}|\leqslant\varepsilon$$

ε 为预给的容差。而迭代初值可取，例如按照向前差公式

$$\boldsymbol{u}_{n+1}^{(0)}=\boldsymbol{u}_n+h\boldsymbol{f}_n$$

或者径直取

$$\boldsymbol{u}_{n+1}^{(0)}=\boldsymbol{u}_n$$

命第 j 次近似 $\boldsymbol{u}_{n+1}^{(j)}$ 与真解 $\boldsymbol{u}_{n+1}$ 的差为 $\boldsymbol{\varepsilon}_{n+1}^{(j)}$

$$\boldsymbol{\varepsilon}_{n+1}^{(j)}=\boldsymbol{u}_{n+1}-\boldsymbol{u}_{n+1}^{(j)}$$

于是(11.5.1)、(11.5.2)相减得

$$\boldsymbol{\varepsilon}_{n+1}^{(j+1)}=h\boldsymbol{f}(\boldsymbol{u}_{n+1},\ t_{n+1})-h\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(j)},\ t_{n+1})\approx h\boldsymbol{A}_{n+1}\boldsymbol{\varepsilon}_{n+1}^{(j)}$$

此处

$$\boldsymbol{A}_{n+1}=\left(\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{u}}\right)_{\boldsymbol{u}_{n+1},t_{n+1}}$$

即

$$\boldsymbol{\varepsilon}_{n+1}^{(j)}=(h\boldsymbol{A}_{n+1})^j\boldsymbol{\varepsilon}_{n+1}^{(0)},\quad j=0,\ 1,\ 2,\ \cdots$$

命 μ_k 为导数矩阵(在线性时就是系数矩阵)的本征值。于是迭代收敛的条件为

$$|h\mu_k|<1,\quad k=1,\ 2,\ \cdots,\ m$$

亦即

$$h<\min_k\frac{1}{|\mu_k|}$$

设本征值 μ_k 为负实数，则时间常数

$$\tau_k=\frac{1}{|\mu_k|}$$

因此收敛的步长条件

$$h<\tau_{\min}\tag{11.5.3}$$

注意迭代收敛条件(11.5.3)与向前差显式的稳定条件(11.4.10)相当。因此，如果采用隐式摆脱稳定条件的限制，以期放大步长，例如在解病态方程时，用比卡迭代法来求解是不行的，因为必须把步长缩回原来的量级。

如果采用隐式改善精度，例如对非病态方程，那时条件(11.5.3)不是问题很大的，比卡迭代法还是可取的，因为它的算法简单。实际计算时可以控制一定的迭代步数；甚至只迭代一次，这就是所谓预估-校正公式。

实践中常常采用一个显式公式作为预估，然后用一个隐式公式作为校正。例如

预估：$\boldsymbol{u}_{n+1}^{(0)}=\boldsymbol{u}_n+h\boldsymbol{f}(\boldsymbol{u}_n,\ t_n)$

校正：$\boldsymbol{u}_{n+1}=\boldsymbol{u}_n+\dfrac{h}{2}[\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(0)},\ t_{n+1})+\boldsymbol{f}(u_n,\ t_n)]$

注意这里向前差预估公式为一阶精度，即

$$\left\{\frac{1}{h}(\boldsymbol{u}_{n+1}-\boldsymbol{u}_n)-\boldsymbol{f}_n\right\}-(\boldsymbol{u}'-\boldsymbol{f})_n=O(h)$$

$$\left\{\frac{1}{h}(\boldsymbol{u}_{n+1}^{(0)}-\boldsymbol{u}_n)-\boldsymbol{f}_n\right\}-(\boldsymbol{u}'-\boldsymbol{f})_n=O(h)$$

于是 $\boldsymbol{u}_{n+1}^{(0)}-\boldsymbol{u}'_{n+1}=O(h^2)$，$\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(0)},\ t_{n+1})-\boldsymbol{f}(\boldsymbol{u}_{n+1},\ t_{n+1})=O(h^2)$，因此预估校正公式(实质是显式)和平均隐式同样达到二阶精度。

一般说来不难验证如果预估显式为 p 阶精度，校正公式为 $p+1$ 阶精度，则迭代一次的效果也达到 $p+1$ 阶精度。因此当预估校正配合得当，在迭代收敛的条件下，迭代的效果主要表现在第一次，而盲目迭代下去则收效不大。这样只迭代一次，即多计算右端函数一次，便能收到提高一阶精度的效果。

附 两种梯形公式

$$\text{I.}\ u_{n+1}=u_n+\frac{h}{2}(f_n+f_{n+1}),\ E=O(h^2),\ \lambda=\frac{1+\frac{1}{2}\mu h}{1-\frac{1}{2}\mu h}$$

$$\text{II.}\ u_{n+1}=u_n+hf\left(\frac{u_n+u_{n+1}}{2},\ t_{n+\frac{1}{2}}\right),\ E=O(h^2),\ \lambda=\frac{1+\frac{1}{2}\mu h}{1-\frac{1}{2}\mu h}$$

两种预估校正

$$\text{I.}\ \begin{cases} u_{n+1}^{(0)}=u_n+hf_n \\ u_{n+1}=u_n+\dfrac{h}{2}(f_n+f_{n+1}^{(0)}) \end{cases}\quad E=O(h^2),\ \lambda=1+\mu h+\frac{1}{2}\mu^2h^2 \tag{11.5.4}$$

即

$$u_{n+1}=u_n+\frac{h}{2}[f_n+f(u_n+hf_n,\ t_{n+1})]$$

$$\text{II.}\ \begin{cases} u_{n+1}^{(0)}=u_n+hf_n \\ u_{n+1}=u_n+hf\left(\dfrac{u_n+u_{n+1}^{(0)}}{2},\ t_{n+\frac{1}{2}}\right) \end{cases}\quad E=O(h^2),\ \lambda=1+\mu h+\frac{1}{2}\mu^2h^2 \tag{11.5.5}$$

即

$$u_{n+1}=u_n+hf\left(u_n+\frac{h}{2}f_n,\ t_{n+\frac{1}{2}}\right)$$

注意这里预估校正实质是显式，但已经不是所谓线性格式(§11.3)，而是非线性格式，是二阶的龙格库塔公式(见§11.8)。

11.5.2 牛顿迭代法与预估校正公式

仍以向后差尤拉隐式(11.3.3)

$$\boldsymbol{u}_{n+1}-\boldsymbol{u}_n=h\boldsymbol{f}(\boldsymbol{u}_{n+1},\ t_{n+1})$$

为例说明牛顿迭代法。设已知 $\boldsymbol{u}_{n+1}^{(j)}$，要求下一级 $\boldsymbol{u}_{n+1}^{(j+1)}$ 基本上满足

$$\boldsymbol{u}_{n+1}^{(j+1)}-\boldsymbol{u}_n=h\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(j+1)},\ t_{n+1})$$

命 $\boldsymbol{u}_{n+1}^{(j+1)}=\boldsymbol{u}_{n+1}^{(j)}+\boldsymbol{\delta}_{n+1}^{(j)}$，于是要求

$$\boldsymbol{u}_{n+1}^{(j)}+\boldsymbol{\delta}_{n+1}^{(j)}-\boldsymbol{u}_n=h\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(j)}+\boldsymbol{\delta}_{n+1}^{(j)},\ t_{n+1})\approx h\boldsymbol{f}_{n+1}^{(j)}+h\boldsymbol{A}_{n+1}^{(j)}\boldsymbol{\delta}_{n+1}^{(j)}$$

此处

$$\boldsymbol{f}_{n+1}^{(j)}=\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(j)},\ t_{n+1})$$

$$\boldsymbol{A}_{n+1}^{(j)}=\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right)(\boldsymbol{u}_{n+1}^{(j)},\ t_{n+1})$$

因此增量 $\boldsymbol{\delta}_{n+1}^{(j)}$ 满足线代数方程组

$$\begin{cases} (\boldsymbol{I}-h\boldsymbol{A}_{n+1}^{(j)})\boldsymbol{\delta}_{n+1}^{(j)}=h\boldsymbol{f}_{n+1}^{(j)}-(\boldsymbol{u}_{n+1}^{(j)}-\boldsymbol{u}_n) \\ \boldsymbol{u}_{n+1}^{(j+1)}=\boldsymbol{u}_{n+1}^{(j)}+\boldsymbol{\delta}_{n+1}^{(j)} \end{cases}\quad j=0,\ 1,\ 2,\ \cdots \tag{11.5.6}$$

直至增量小于预给的容差为止：

$$|\boldsymbol{\delta}_{n+1}^{(j)}|\leqslant\varepsilon$$

初始值则可以取，例如

$$\boldsymbol{u}_{n+1}^{(0)}=\boldsymbol{u}_n$$

注意，如果原微分方程组是线性的，即 $\boldsymbol{u}'=\boldsymbol{f}(\boldsymbol{u},\ t)=\boldsymbol{A}(t)\boldsymbol{u}+\boldsymbol{g}(t)$ 则成为

$$\boldsymbol{u}_{n+1}-\boldsymbol{u}_n=h\boldsymbol{A}_{n+1}\boldsymbol{u}_{n+1}+h\boldsymbol{g}_{n+1} \tag{11.5.7}$$

注意由于方程组是线性的，导数矩阵就等于系数矩阵，即

$$\boldsymbol{A}_{n+1}^{(j)}=\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right)_{\boldsymbol{u}_{n+1}^{(j)},\,t_{n+1}}=\boldsymbol{A}(t_{n+1})=\boldsymbol{A}_{n+1}$$

与 j 无关。

当取初值 $\boldsymbol{u}_{n+1}^{(0)}=\boldsymbol{u}_n$ 时，方程(11.5.7)本身与它的牛顿线性化方程(11.5.6)相一致，即牛顿迭代一步就是真解 $\boldsymbol{u}_{n+1}^{(1)}=\boldsymbol{u}_{n+1}$，即使矩阵 $\boldsymbol{A}_{n+1}$ 高度病态也是如此。在非线性的情况下，只要初始值靠近真解(实践上总能作到的)，总是保证收敛的。这就不像比卡迭代那样步长要受最小时间常数(即最大本征值)的严重限制。这是牛顿迭代的有利之点。因此它特别适应于病态方程。

比卡方法不论对线性或非线性方程都按统一而简单算法执行，而牛顿法每迭代一步需要解一个线代数方程组，这是不利之点，还留有解法的问题，解法的选取适当与否关系到牛顿法的成败。

线代数方程最好采取直接法如消去法。设未知数个数为 m，如视矩阵 $\boldsymbol{A}$ 为满矩阵，当 m 很大时，存储量 $\sim m^2$ 及工作量 $\sim m^3$ 都是浩大的。但是在实际上求解的微分方程的导数阵 $\boldsymbol{A}$(特别对于病态方程而言)几乎都是稀疏矩阵，即零元素占压倒多数，而且矩阵的非零结构在积分过程中不变，这是有利的因素。因此一般采取稀疏矩阵的技术可以争取到解一次方程的工作量线性地依赖于 m。

此外，在病态方程通常还有这样的特点，即导数矩阵 $\boldsymbol{A}$ 是缓变的。因此在牛顿迭代过程无须每一个迭代步都对 $\boldsymbol{A}_{n+1}^{(j)}$ 进行更新，甚至于也无须在每一个积分步进行更新。这样可以节约产生导数矩阵的计算量。

牛顿迭代一次的预估校正方法:

类似于比卡迭代一次方法，也可以构造牛顿迭代一次的方法。鉴于在线性方程时牛顿法迭代一次就已达到真解，因此可以设想牛顿迭代一次的方法可以保持隐式恒稳的优点。以向后差隐式为例，取

外插预估
$$\boldsymbol{u}_{n+1}^{(0)}=\boldsymbol{u}_n \tag{11.5.8}$$

隐式校正
$$\begin{cases}(\boldsymbol{I}-h\boldsymbol{A}_{n+1})\boldsymbol{\delta}_{n+1}=h\boldsymbol{f}_{n+1}^{(0)}\\ \boldsymbol{u}_{n+1}=\boldsymbol{u}_{n+1}^{(0)}+\boldsymbol{\delta}_{n+1}\end{cases} \tag{11.5.9}$$

此处
$$\boldsymbol{A}_{n+1}=\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right)_{\boldsymbol{u}_{n+1}^{(0)},\,t_{n+1}}=\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right)_{\boldsymbol{u}_n,\,t_{n+1}}$$

$$\boldsymbol{f}_{n+1}^{(0)}=\boldsymbol{f}(\boldsymbol{u}_{n+1}^{(0)},\ t_{n+1})=\boldsymbol{f}(\boldsymbol{u}_n,\ t_{n+1})$$

它事实上就是以(11.5.8)为初值，牛顿迭代一次。也可以表为如下的隐式格式

$$\frac{1}{h}(\boldsymbol{u}_{n+1}-\boldsymbol{u}_n)=\boldsymbol{A}_{n+1}\boldsymbol{u}_{n+1}-\boldsymbol{A}_{n+1}\boldsymbol{u}_n+\boldsymbol{f}(\boldsymbol{u}_n,\ t_{n+1}) \tag{11.5.10}$$

由于 $\boldsymbol{u}_{n+1}=\boldsymbol{u}_n+O(h)$，因此不难验证(11.5.10)和向后差隐式(11.5.1)同样是一阶精度，即 $E=O(h)$。

当 $\boldsymbol{u}'=\boldsymbol{f}(\boldsymbol{u},\ t)=\boldsymbol{A}(t)\boldsymbol{u}+\boldsymbol{g}(t)$ 即微分方程为线性时

$$\boldsymbol{f}(\boldsymbol{u}_n,\ t_{n+1})=\boldsymbol{A}_{n+1}\boldsymbol{u}_n+\boldsymbol{g}_{n+1}$$

$$\boldsymbol{f}(\boldsymbol{u}_{n+1},\ t_{n+1})=\boldsymbol{A}_{n+1}\boldsymbol{u}_{n+1}+\boldsymbol{g}_{n+1}$$

因此差分格式(11.5.10)与(11.5.1)等同。因此线性化分析的意义下两者的稳定性能相同，即为恒稳的。

§ 11.6 基于数值积分的方法

(§ 11.3)中所介绍的尤拉公式，精度为一阶或二阶，是低精度的公式，比较适合于解或其导数有间断的情况。当解具有高光滑度时，它们的效率不高，这时必须利用高精度的差分公式。它们可以分为三类，即基于数值积分的方法(§11.6)，基于数值微分的方法(§11.7)和基于幂次展开的方法(§11.8)。前二者都运用多项式插值的思想，都是线性多步格式，第三类则是非线性单步格式即龙格库塔法。

微分方程

$$u'=f(u(t),\ t) \tag{11.6.1}$$

可以写成积分的形式

$$u_{n+1}-u_n=\int_{t_n}^{t_{n+1}} f(u(t),\ t)dt$$

或者写成

$$\frac{1}{h}(u_{n+1}-u_n)=\frac{1}{h}\int_{t_n}^{t_{n+1}} f(u(t),\ t)dt,\quad t_{n+1}=t_n+h \tag{11.6.2}$$

这是准确的公式。

现用 k 个节点 $t_n,\ t_{n-1},\ \cdots,\ t_{n-k+1}$ 的 f 值 $f_n,\ f_{n-1},\ \cdots,\ f_{n-k+1}$ 插出一个 $k-1$ 次多项式 $F(t)$

$$f(u(t),\ t)\approx F(t)=f_n l_n(t)+f_{n-1}l_{n-1}(t)+\cdots+f_{n-k+1}l_{n-k+1}(t) \tag{11.6.3}$$

这里 $l_{n-j}(t)$ 就是插值的基函数，即在节点 t_{n-j} 取值为1在其它节点取值为0的 $k-1$ 次多项式。以(11.6 3)代入(11.6.2)的右端，实际上是外插到区间 $[t_n,\ t_{n+1}]$ 上积分(图11 6)，就得到显式亚丹斯公式

$$\frac{1}{h}(u_{n+1}-u_n)=\beta_1 f_n+\beta_2 f_{n-1}+\cdots+\beta_k f_{n-k+1} \tag{11.6.4}$$

$$\beta_j=\frac{1}{h}\int_{t_n}^{t_{n+1}} l_{n-j+1}(t)dt,\quad j=1,\ 2,\ \cdots,\ k \tag{11.6.5}$$

在等距节点即 $t_{n-j}=t_n-jh$ 时系数 β_j 见表 11.1 左。当 $k=1$ 时就是尤拉向前差公式。

也可以用 k 个节点 $t_{n+1},\ t_n,\ \cdots,\ t_{n-k+2}$ 的值 $f_{n+1},\ f_n,\ \cdots,\ f_{n-k+2}$ 插出一个 $k-1$ 次多项式 $F(t)$

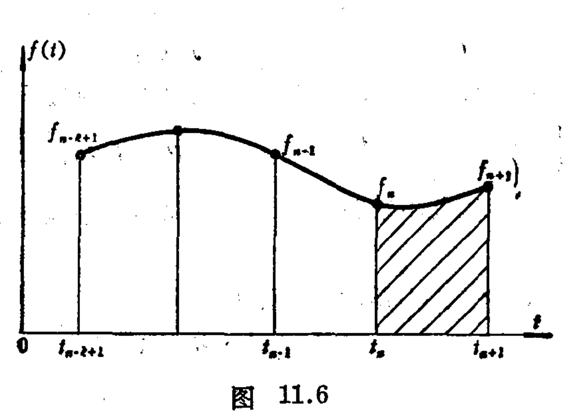


图 11.6

$$f(u(t),\ t)\approx F(t)=f_{n+1}l_{n+1}(t)+f_n l_n(t)+\cdots+f_{n-k+2}l_{n-k+2}(t) \tag{11.6.6}$$

$l_{n+1-j}(t)$是在节点t_{n+1-j}取值为1在其它节点取值为0的$k-1$次多项式。以(11.6.6)代入(11.6.2)的右端,注意这时是内插,在区间$[t_n,\ t_{n+1}]$上积分,于是得到隐式亚丹斯公式

$$\frac{1}{h}(u_{n+1}-u_n)=\beta_0 f_{n+1}+\beta_1 f_n+\cdots+\beta_{k-1}f_{n-k+2} \tag{11.6.7}$$

$$\beta_j=\frac{1}{h}\int_{t_n}^{t_{n+1}} l_{n+1-j}(t)\,dt,\quad j=0,\ 1,\ \cdots,\ k-1 \tag{11.6.8}$$

在等距节点时系数β_j见表11.1右。当$k=0,\ 1$时分别就是尤拉向后差及平均公式。

表11.1 亚丹斯公式系数表

	显式						隐式					
k	β_1	β_2	β_3	β_4	β_5	β_6	β_0	β_1	β_2	β_3	β_4	β_5
0	1						1					
1	$\frac{3}{2}$	$-\frac{1}{2}$					$\frac{1}{2}$	$\frac{1}{2}$				
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$				$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$			
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$			$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$		
4	$\frac{1901}{720}$	$-\frac{2744}{720}$	$\frac{2616}{720}$	$-\frac{1274}{720}$	$\frac{251}{720}$		$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$	
5	$\frac{4727}{1440}$	$-\frac{7673}{1440}$	$\frac{9182}{1440}$	$-\frac{6798}{1440}$	$\frac{2627}{1440}$	$-\frac{425}{1440}$	$\frac{475}{1440}$	$\frac{1427}{1440}$	$-\frac{798}{1440}$	$\frac{482}{1440}$	$-\frac{173}{1440}$	$\frac{27}{1440}$

截断误差和稳定性

在亚丹斯公式中,取了k个节点的f值插出$k-1$次多项式,根据第一章1.2.2节,有误差估计

$$f(u(t),\ t)-F(t)=O(h^k),\quad t_n\leqslant t\leqslant t_{n+1}$$

隐式对于显式的差别只在于估计式中h^k前面的系数要小一些,再作积分,则有

$$\int_{t_n}^{t_{n+1}} f(u(t),\ t)\,dt-\int_{t_n}^{t_{n+1}} F(t)\,dt=O(h^{k+1})$$

$$\frac{1}{h}\int_{t_n}^{t_{n+1}} f(u(t),\ t)\,dt-\frac{1}{h}\int_{t_n}^{t_{n+1}} F(t)\,dt=O(h^k)$$

因此亚丹斯公式(11.6.4)和(11.6.7)的截断误差都是$E=O(h^k)$,即具有k阶精度。

注意:当右项$f\equiv1$时,亚丹斯公式应准确成立,因此恒有

$$\beta_1+\beta_2+\cdots+\beta_k=1(\text{显式})$$
$$\beta_0+\beta_1+\cdots+\beta_{k-1}=1(\text{隐式})$$
(1

亚丹斯公式(11.6.4)、(11.6.7)的本征方程统一写为

$$\lambda^m-\lambda^{m-1}=z(\beta_0\lambda^m+\beta_1\lambda^{m-1}+\cdots+\beta_m),\ z\equiv\mu h$$

k 阶精度显式时 $m=k$, $\beta_0=0$; k 阶精度隐式时 $m=k-1$。视 z 为小参数。当 $z=0$ 时方程 $\lambda^m-\lambda^{m-1}=0$, 即有单根 $\lambda=1$ 及 $m-1$ 重根 $\lambda=0$; 当 $z\sim0$ 时根 $\lambda=0$ 受微扰而仍 ~0, 不影响稳定性; 根 $\lambda=1$ 的受微扰而成为

$$\lambda(z)=1+az+O(z^2)$$

代入(11.6.10), 比较 z 的系数并利用关系(11.6.9)可知 $a=1$ 即

$$\lambda(\mu h)=1+\mu h+O(h^2)$$

$$|\lambda(\mu h)|^2=1+2\mathrm{Re}\mu h+O(h^2)$$

因此当 $\mathrm{Re}\mu<0$ 而 h 充分小时恒有 $|\lambda(\mu h)|<1$, 而步长 h 限制在时间常数

$$\tau=\frac{1}{|\mathrm{Re}\mu|}$$

的量级, 因此亚丹斯公式不论何阶, 或显或隐, 都至少是条件稳的。如果按(11.6.10)

$$\mu h\equiv z=\frac{\lambda^m-\lambda^{m-1}}{\beta_0\lambda^m+\cdots+\beta_m}$$

画出 z 平面上的临界曲线(对应于 $|\lambda|=1$ 即 $\lambda=e^{i\theta}$)则大致如图 11.7(显式)及图 11.8(隐式)。当 k 增大时稳定域缩小。除了隐式 $k=1$, $k=2$ 为恒稳尤拉公式外, 其它显隐各式的稳定域都是左半平面内的有限域, 即仅仅是条件稳。

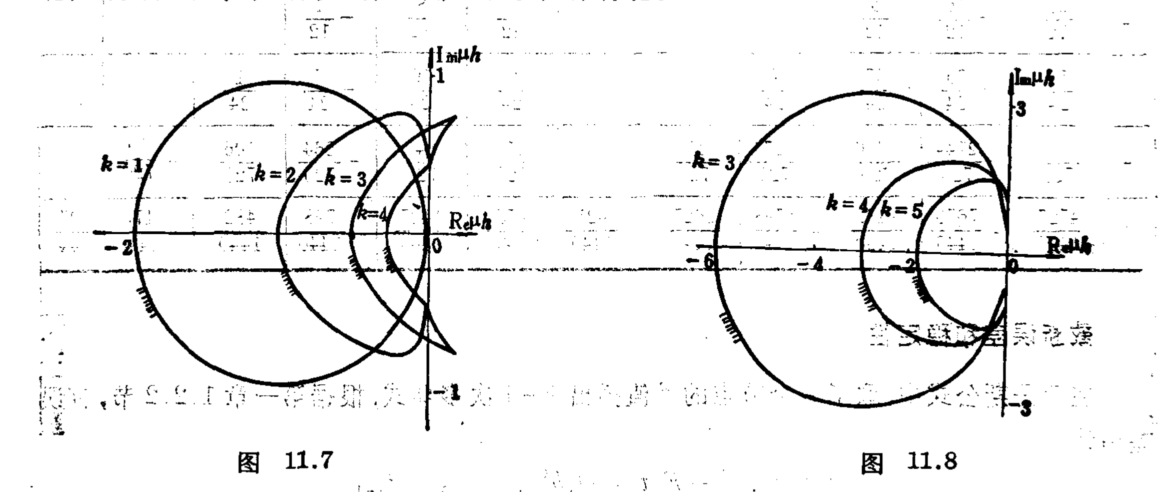


图 11.7　　　　图 11.8

常用的亚丹斯公式是

(1) 四阶精度的显式

$$u_{n+1}=u_n+\frac{h}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3}),\quad E=O(h^4)\tag{11.6.11}$$

(2) 四阶精度的预估校正式

$$\left.\begin{aligned}u_{n+1}^{(0)}&=u_n+\frac{h}{12}(23f_n-16f_{n-1}+5f_{n-2})\\u_{n+1}&=u_n+\frac{h}{24}(9f_{n+1}^{(0)}+19f_n-5f_{n-1}+f_{n-2})\end{aligned}\right\}\tag{11.6.12}$$

青度为 $O(h^3)$, 迭代一次, 精度仍达 $O(h^4)$ (见 §11.5)。

斯公式适用于有光滑解的非病态方程。每步只须计算 f 一次(用预估校正时为两

较小。注意即使是隐式, 除了 $k=0$, $k=1$ 以外, 也不适于解病态方程。

§ 11.7 基于数值微分的方法

在这里不从积分形式(11.6.2)而直接从微分形式(11.6.1)

$$u'(t)=f(u(t),\ t) \tag{11.7.1}$$

出发。用 $k+1$ 个节点 $t_{n+1},\ t_n,\ \cdots,\ t_{n-k+1}$ 的 u 值 $u_{n+1},\ u_n,\ \cdots,\ u_{n-k+1}$ 插出一个 k 次多项式 $U(t)$:

$$u(t)\approx U(t)=u_{n+1}l_{n+1}(t)+u_nl_n(t)+\cdots+u_{n-k+1}l_{n-k+1}(t) \tag{11.7.2}$$

设在 $t=t_n$ 处微分(图 11.9), 即 $u'(t_n)\approx U'(t_n)$ 即得显式公式

$$\frac{1}{h}(\alpha_0u_{n+1}+\alpha_1u_n+\cdots+\alpha_ku_{n-k+1})=f_n \tag{11.7.3}$$

$$\alpha_j=hl'_{n+1-j}(t_n) \tag{11.7.4}$$

取等距节点,

当 $k=1$ 时, $\alpha_0=1,\ \alpha_1=-1$, 就是尤拉向前差公式;

当 $k=2$ 时, $\alpha_0=\frac{1}{2},\ \alpha_1=0,\ \alpha_2=-\frac{1}{2}$, 则是中心差公式, 恒不稳(11.3.1 节、11.4.2 节);

当 $k=3$ 时, $\alpha_0=\frac{1}{3},\ \ \alpha_1=\frac{1}{2},\ \ \alpha_2=-1,\ \ \alpha_3=\frac{1}{6},\ \cdots$。

可以证明, 当 $k\geqslant3$ 时也都是恒不稳的, 因此这是一个基本无用的序列。

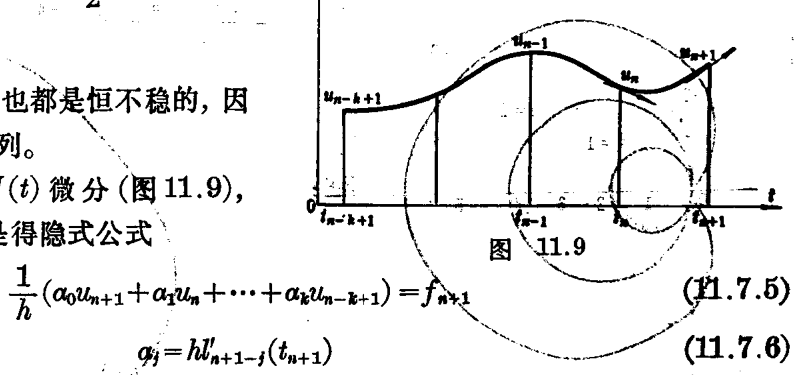


图 11.9

设改在 $t=t_{n+1}$ 处对 $U(t)$ 微分(图 11.9), 即 $u'(t_{n+1})\approx U'(t_{n+1})$, 于是得隐式公式

$$\frac{1}{h}(\alpha_0u_{n+1}+\alpha_1u_n+\cdots+\alpha_ku_{n-k+1})=f_{n+1} \tag{11.7.5}$$

$$\alpha_j=hl'_{n+1-j}(t_{n+1}) \tag{11.7.6}$$

在等距节点时, 系数 α_j 见表 11.2。当 $k=1$ 时就是尤拉向后差公式。下面可以看到, 与显式(11.7.3)不同, 隐式(11.7.5)是一个相当有用的序列。

表 11.2 基于数值微分的隐式系数

k	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	1	-1					
2	$\frac{3}{2}$	$-\frac{4}{2}$	$\frac{1}{2}$				
3	$\frac{11}{6}$	$-\frac{18}{6}$	$\frac{9}{6}$	$\frac{2}{6}$			
4	$\frac{25}{12}$	$-\frac{48}{25}$	$\frac{36}{12}$	$-\frac{16}{12}$	$\frac{3}{12}$		
5	$\frac{137}{60}$	$-\frac{300}{60}$	$\frac{300}{60}$	$-\frac{200}{60}$	$\frac{75}{60}$	$-\frac{12}{60}$	
6	$\frac{147}{60}$	$-\frac{360}{60}$	$\frac{450}{60}$	$-\frac{400}{60}$	$\frac{225}{60}$	$-\frac{72}{60}$	$\frac{10}{60}$

根据§1.2，$k+1$个点k次插值的一阶导数的误差(在节点处)为$u'-U'=O(h^k)$，因此公式(11.7.5)以及(11.7.3)的截断误差是$O(h^k)$，即具有k阶精度。

稳定性

差分格式(11.7.5)的本征方程是

$$\mu h \equiv z = \frac{\alpha_0\lambda^k + \alpha_1\lambda^{k-1} + \cdots + \alpha_k}{\lambda^k} \equiv f(\lambda) \tag{11.7.7}$$

图11.10表示$k=1, 2, 3$的临界曲线，其外部为稳定域。图11.11表示了$k=1-6$的临界曲线，每根曲线的左方为稳定域，由于感兴趣的左半平面$\mathrm{Re}\,\mu h \leqslant 0$，并由于上下对称故只画第三象限。$k=1, 2$的稳定域包含了左半平面，因此是恒稳的。$k=3, 4, 5, 6$虽然不是恒稳，但近于恒稳，即其稳定域均向左展至无穷，并在靠近虚轴处包含一条包住负实轴的"走廊"(图11.12)。具体地说有下列两个性质：

1° 当$\mathrm{Re}\,\mu h \leqslant -6.1$时稳定。

2° 当$-6.1 \leqslant \mathrm{Re}\,\mu h \leqslant 0$时在$|I_m \mu h| \leqslant 0.5$内为稳定。当$k>6$以后不再有此性质。

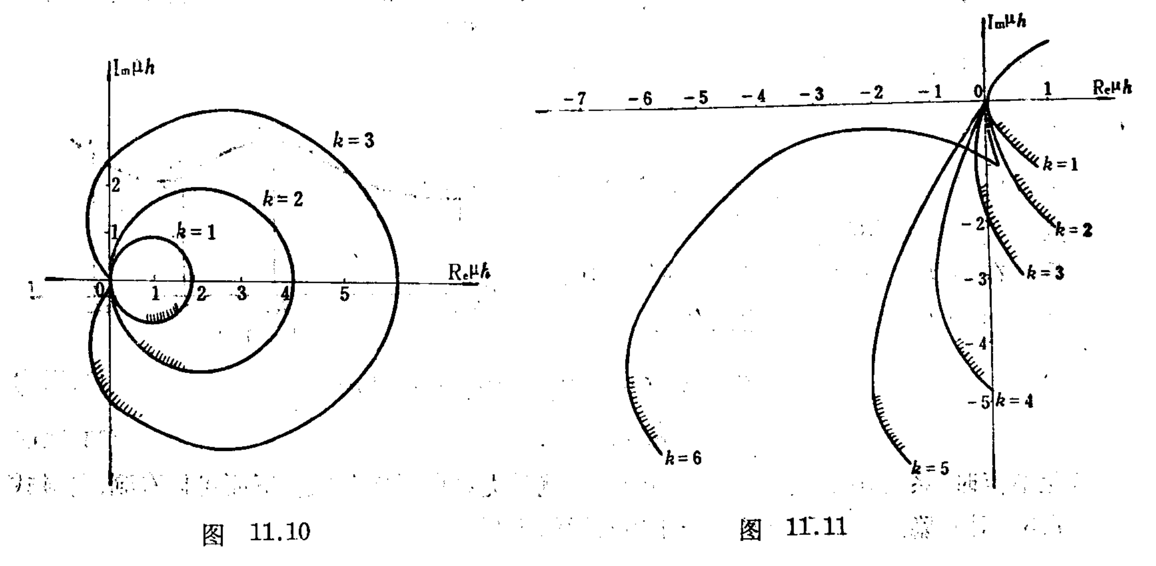


图 11.10　　　　图 11.11

我们知道(§11.4)恒稳格式如尤拉向后差和梯形公式是适合解病态方程的。可以证明(见[2])恒稳的线性多步格式

$$\alpha_0 u_{n+1} + \cdots + \alpha_k u_{n+1-k} = h(\beta_0 f_0 + \cdots + \beta_k f_{n+1-k})$$

中最高只能达到二阶精度，在这个意义下梯形公式是最优的。但是这并不排除在非线性格式(如§11.8)中有可能构造高于二阶的恒稳公式。另一方面，对于病态方程也不一定要求恒稳。事实上病态方程的本征值在左半平面内向$-\infty$散得很远，于是上面所述的性质1°可以适应。如果有本征根靠虚轴，即$\mathrm{Re}\,\mu \sim 0$而虚部的模$|\beta| = |I_m \mu|$很大，这对应于以$\frac{\beta}{2\pi}$为频率(即以$\frac{2\pi}{\beta}$为周期，见11.1.2节的高频低阻尼的周期振动)。从截差来考虑，从直观上看，为了保证必要的精度，一般应该在谐波的一个周期内有大约12个节点，即取

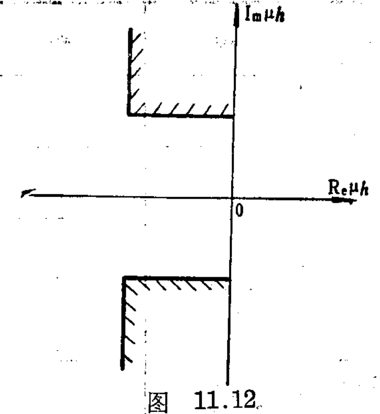


图 11.12

$$h \leqslant \frac{2\pi}{12|\beta|} \approx \frac{1}{2|\beta|} \quad \text{即} \quad |I_m \mu h| \leqslant \frac{1}{2} \tag{11.7.8}$$

因此即使格式是恒稳的，步长仍须受此制约，而上面所述的性质 2° 正好适合于此。因此格式(11.7.5)($k=1-6$)的近于恒稳的性质对于病态方程已经足够，而它有高达六阶的精度。因此这是适于病态方程的高精度方法，它也正是由于解病态方程的需要而被重视的[3, 4]。

在 11.5.2 节所述牛顿迭代法也可以推广到这里来解隐式方程。

§ 11.8 基于幂级数展开的方法

推广尤拉方法还有另外的途径，即用幂级数展开的方法导出高精度的单步公式，即所谓龙格-库塔(Runge-Kutta)方法。

考虑一个微分方程

$$u' = f(u(t),\ t) \tag{11.8.1}$$

从积分公式

$$(u_{n+1} - u_n) = \int_{t_n}^{t_n+h} f(u(t),\ t)\,dt \tag{11.8.2}$$

出发，将积分号下函数对 $t=t_n$ 点作幂次展开，命 f' 为 $f(u(t),\ t)$ 对 t 的全微商

$$f' = f_u \cdot u' + f_t = f \cdot f_u + f_t = \left(f\frac{\partial}{\partial u} + \frac{\partial}{\partial t}\right) f$$

$$f'' = \left(f\frac{\partial}{\partial u} + \frac{\partial}{\partial t}\right)^2 f$$

$$f''' = \left(f\frac{\partial}{\partial u} + \frac{\partial}{\partial t}\right)^3 f$$

$$\cdots\cdots$$

于是

$$f(u(t),\ t) = \sum_{m=0}^{k-1} f_{t_n}^{(m)} \cdot \frac{(t-t_n)^m}{m!} + O(h^k)$$

$$\int_{t_n}^{t_n+h} f(u(t),\ t)\,dt = h \sum_{m=0}^{k-1} \left(\left(f\frac{\partial}{\partial u} + \frac{\partial}{\partial t}\right)^m f\right)_{t_n} \frac{h^m}{(m+1)!} + O(h^{k+1}) \tag{11.8.3}$$

这里要用到 $t=t_n$ 处 f 的各阶偏导数，需要解析计算，这是不方便的。龙格-库塔方法的要点在于计算从 t_n 起适当的 k 个点 f 值作为代替而可达到同阶的精度，命

$$\int_{t_n}^{t_n+h} f(u(t),\ t)\,dt = \sum_{i=1}^{k} a_i \varphi_i + O(h^k) \tag{11.8.4}$$

$$\left.\begin{aligned}
&\varphi_1 = hf(u_n,\ t_n)\\
&\varphi_2 = hf(u_n + b_{2,1}\varphi_1,\ t_n + c_2 h)\\
&\varphi_3 = hf(u_n + b_{3,1}\varphi_1 + b_{3,2}\varphi_2,\ t_n + c_3 h)\\
&\cdots\cdots\\
&\varphi_k = hf(u_n + b_{k,1}\varphi_1 + \cdots + b_{k,k-1}\varphi_{k-1},\ t_n + c_k h)
\end{aligned}\right\} \tag{11.8.5}$$

将(11.8.4)的右端在 $t=t_n$ 点作幂次展开并与(11.8.3)右端展式比较对应幂次和对应阶偏导数，就可以定出诸系数 $a,\ b,\ c$。一般说来，结果是不唯一的，这样就可以得到不同种类的龙格-库塔公式

$$\frac{1}{h}(u_{n+1} - u_n) = \frac{1}{h}\sum_{i=1}^{k} a_i \varphi_i, \quad \text{截差 } E = O(h^k) \tag{11.8.6}$$

对 $k=1$–4 各举两种见表 11.3。这些系数同样适用于一个方程或方程组。

表 11.3 龙格-库塔公式表

$k=1$	$\varphi_1=hf(u_n, t_n)$ $u_{n+1}=u_n+\varphi_1$	
$k=2$	$\varphi_1=hf(u_n, t_n)$ $\varphi_2=hf\left(u_n+\frac{1}{2}\varphi_1, t_n+\frac{1}{2}h\right)$ $u_{n+1}=u_n+\varphi_2$	$\varphi_1=hf(u_n, t_n)$ $\varphi_2=hf(u_n+\varphi_1, t_n+h)$ $u_{n+1}=u_n+\frac{1}{2}(\varphi_1+\varphi_2)$
$k=3$	$\varphi_1=hf(u_n, t_n)$ $\varphi_2=hf\left(u_n+\frac{1}{3}\varphi_1, t_n+\frac{1}{3}h\right)$ $\varphi_3=hf\left(u_n+\frac{2}{3}\varphi_2, t_n+\frac{2}{3}h\right)$ $u_{n+1}=u_n+\frac{1}{4}(\varphi_1+0\cdot\varphi_2+3\varphi_3)$	$\varphi_1=hf(u_n, t_n)$ $\varphi_2=hf\left(u_n+\frac{1}{2}\varphi_1, t_n+\frac{1}{2}h\right)$ $\varphi_3=hf(u_n-\varphi_1+2\varphi_2, t_n+h)$ $u_{n+1}=u_n+\frac{1}{6}(\varphi_1+4\varphi_2+\varphi_3)$
$k=4$	$\varphi_1=hf(u_n, t_n)$ $\varphi_2=hf\left(u_n+\frac{1}{2}\varphi_1, t_n+\frac{1}{2}h\right)$ $\varphi_3=hf\left(u_n+\frac{1}{2}\varphi_2, t_n+\frac{1}{2}h\right)$ $\varphi_4=hf(u_n+\varphi_3, t_n+h)$ $u_{n+1}=u_n+\frac{1}{6}(\varphi_1+2\varphi_2+2\varphi_3+\varphi_4)$	$\varphi_1=hf(u_n, t_n)$ $\varphi_2=hf\left(u_n+\frac{1}{3}\varphi_1, t_n+\frac{1}{3}h\right)$ $\varphi_3=hf\left(u_n-\frac{1}{3}\varphi_1+\varphi_2, t_n+\frac{2}{3}h\right)$ $\varphi_4=hf(u_n+\varphi_1-\varphi_2+\varphi_3, t_n+h)$ $u_{n+1}=u_n+\frac{1}{8}(\varphi_1+3\varphi_2+3\varphi_3+\varphi_4)$

这里 $k=1$ 就是尤拉向前差公式，$k=2$ 两种都是以两种尤拉梯形公式为基础的预估校正公式(11.5.4)、(11.5.5)。龙格-库塔公式都是单步显式。k 阶精度的公式每步需要计算右端函数 k 次。$k\geqslant 2$ 以后都是非线性的差分格式。最常用的是 $k=4$。

稳定性

为了判稳，考虑模型方程

$$u'=f(u, t)=\mu u$$

我们无需将 $f=\mu u$ 代入各个具体的格式而只须利用(11.8.2)、(11.8.4)将 $f=\mu u_n$ 代入(11.8.3)的右端幂次展开即得

$$u_{n+1}-u_n=\left(\sum_{m=0}^{k-1}\frac{(\mu h)^{m+1}}{(m+1)!}\right)u_n \tag{11.8.7}$$

因此本征方程为

$$\lambda=1+\mu h+\frac{1}{2!}(\mu h)^2+\cdots+\frac{1}{k!}(\mu h)^k, \quad \mu h=z \tag{11.8.8}$$

这就是 $e^{\mu h}$ 的幂级数，截取到 k 次。

$$|\lambda|^2=1+2\operatorname{Re}\mu h+O(h^2)$$

因此对 $\operatorname{Re}\mu<0$，当 h 充分小时 $|\lambda|<1$，而步长限制在时间常数

$$\tau=\frac{1}{|\operatorname{Re}\mu|}$$

的量级，因此龙格-库塔公式是条件稳的。

当 $\operatorname{Re}\mu=0$，即 $\mu=i\beta$，则以常用 4 阶公式为例可以算出

$$|\lambda|^2=\begin{cases}1+(\beta h)^2=1+O(h^2), & k=1\\ 1+\dfrac{1}{4}(\beta h)^4=1+O(h^4), & k=2\\ 1-\dfrac{1}{12}(\beta h)^4+\dfrac{1}{36}(\beta h)^6=1+O(h^4), & k=3\\ 1-\dfrac{1}{12}(\beta h)^6+\dfrac{1}{(24)^2}(\beta h)^8=1+O(h^6), & k=4\end{cases}$$

因此基本能保持临界稳定性，特别当 $k=4$。

在 $z=\mu h$ 平面上的临界曲线（对应于 $|\lambda|=1$）和稳定域示意如图 11.13。$k=1$ 时就是中心在 $z=-1$ 的单位圆，k 增大时稳定域（都是有限域）逐次扩大，$k\to\infty$ 的极限就是左半平面。由于它们仅是条件稳，因此不适于病态方程。为了病态问题的需要也可以构造恒稳的隐式龙格-库塔公式。

图 11.13

§11.9 方法概述

常微数值解法中用得较广的是四阶龙格-库塔方法。它除了具有较高精度外，由于是单步的，故有自动起步和便于改变步长的优点。对于稳定的系统，只要步长充分小就能保证数值稳定性，而且也能近似地保持临界稳定性。它的缺点是每步需计算右端四次，工作量较大；但是例如在预估校正型方法中也要计算两次或更多，因此这并不是严重的缺点。由于显式龙格-库塔法的步长限制在最小时间常数的量级，因此不适于解病态方程。

病态微分方程在计算实践中愈来愈占重要的地位，包括龙格-库塔方法在内的一般经典显式方法是不适应的。对此应采用恒稳或近于恒稳的格式，例如尤拉向后差及梯形公式以及高精度的基于数值微分的隐式格式。目前在隐式求解时采用牛顿迭代，并相应地利用矩阵的稀疏结构的特点用直接法解线代数方程，这类问题的方法正在发展中，见[3]，[5]。

对于解不光滑如函数或导数有间断的情况一般宁愿采用低精度但比较安全可靠的方法，特别是几种尤拉公式。

另有一种低阻尼的简振运动和轨道方程问题，解具有极高的光滑度，而且方程是纯二阶的，并具有近于临界稳定的特点。一般针对二阶方程形式制定格式如考埃尔(Cowell)方法等，本章未涉及，可参看[1]。

此外，关于算法执行中自动调整步长和自动变阶的问题，在实践上是很重要的，本章也未涉及，可以参看[3，4]。

参考资料

[1] Henrici, "Discrete Variable Methods in Ordinary Differential Equations", New York, 1962.
[2] Dahlquist, "A special stability problem for linear multistep method", BIT V. 3(1963) pp 27~43.
[3] Gear, "Automatic integration of stiff ordinary differential equations", Information Processing 1968 V. 1 p 187.
[4] Gear, "Numerical Solution Initial Value Problems in Ordinary Differential Equations", New York, 1971.
[5] Willougby, ed., "Stiff Differential Systems, An International Symposium", New York, 1973.

第十二章　偏微分方程初值问题数值解法

§12.1　几个典型方程的特点

连续介质的动态物理过程通常表为带时间 t 的偏微分方程。最简单的典型形式有：

对流方程(双曲型)

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0 \tag{12.1.1}$$

扩散方程(抛物型)

$$\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}=0,\quad b>0 \tag{12.1.2}$$

以及两者相耦合的对流-扩散方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}-b\frac{\partial^2 u}{\partial x^2}=0 \tag{12.1.3}$$

设给定初值条件

$$u(x,0)=\varphi(x),\quad -\infty<x<\infty \tag{12.1.4}$$

很容易验证，对流方程(12.1.1)的解为

$$u(x,t)=\varphi(x-at),\quad -\infty<x<\infty,\quad t\geqslant 0 \tag{12.1.5}$$

这意味着在 $x-t$ 平面上沿每根直线 $x-at=\text{const}$ u 值保持不变。这种直线叫做特征线(图 12.1 及 12.2)。如果把(12.1.4)看作初始时刻 $t=0$ 的波形，则(12.1.5)表示这个波以速度 $|a|$ 传播，当 $a>0$ 时沿 x 方向传播，当 $a<0$ 时沿 $-x$ 方向传播，而波形保持不变。图 12.3 显示了初始为三角波形的演化过程。在更一般的双曲型方程时，波的形状、幅度等均可有变化，但是扰动恒以有限的速度传播，并能保持明确的波阵面。这是波动过程的共同特点，在数学上表为双曲型方程。对流方程(12.1.1)表示一个单向传播的波，也叫做单向波方程。

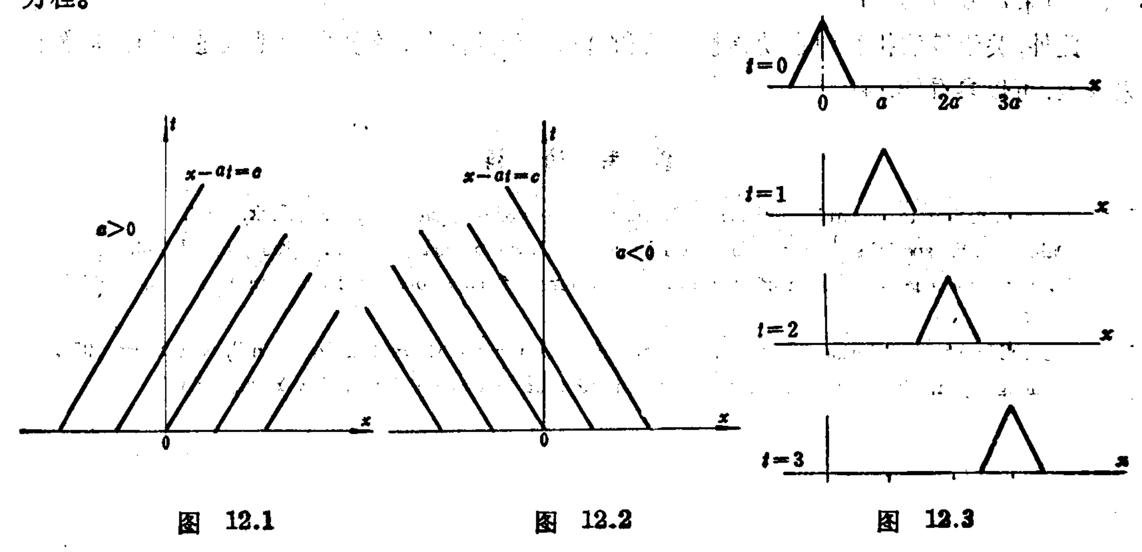


图 12.1　　图 12.2　　图 12.3

在物理上常见的波动方程

$$\frac{\partial^2 w}{\partial t^2}-a^2\frac{\partial^2 w}{\partial x^2}=0 \tag{12.1.6}$$

可以化为联立的单向波方程组的形式。例如，命 $v_1=\frac{\partial w}{\partial t}$，$v_2=a\frac{\partial w}{\partial x}$，则(12.1.6)就可化为一阶微分方程组

$$\begin{cases}\frac{\partial v_1}{\partial t}-a\frac{\partial v_2}{\partial x}=0\\ \frac{\partial v_2}{\partial t}+a\frac{\partial v_1}{\partial x}=0\end{cases} \tag{12.1.7}$$

再命 $u_1=\frac{1}{\sqrt{2}}(v_1-v_2)$，$u_2=\frac{1}{\sqrt{2}}(v_1+v_2)$ 则(12.1.7)又化为

$$\begin{cases}\frac{\partial u_1}{\partial t}-a\frac{\partial u_1}{\partial x}=0\\ \frac{\partial u_2}{\partial t}+a\frac{\partial u_2}{\partial x}=0\end{cases} \tag{12.1.8}$$

这表示两个沿相反方向传播的单向波。因此，单向波方程虽然简单，但有很大的代表性。

扩散方程(12.1.2)在初值条件(12.1.4)下的解可以表为

$$u(x,t)=\frac{1}{\sqrt{4\pi bt}}\int_{-\infty}^{\infty}\exp\left[-\frac{(x-\xi)^2}{4bt}\right]\varphi(\xi)d\xi,\quad -\infty<x<\infty,\quad t\geqslant 0 \tag{12.1.9}$$

图 12.4 给出了初始的三角形分布以及以后各时刻的演化。特点是波形的棱角消失，逐渐平滑化。物质的弥散过程和热传导中的温度等化过程都是这样的。不管初始分布如何集中，它总是在瞬刻之间影响于无穷，可以说是以无限的速度来传播影响的，虽然这种影响是随距离按指数状衰减。这是一切弛豫过程的共同特点，在数学上表为抛物型方程。

设介质的长度为 L。一个不均匀的初始分布在弛豫过程中将趋于等化。可以用量纲分析来估计等化所需的时间 T。显然 T 不依赖于分布 u 本身而只依赖于问题的其它参数即扩散系数 b(量纲为(长度)^2/时间)和 L(量纲为长度)。b 与 L 只有一个组合即 L^2/b 具有时间的量纲，因此在量级上应有

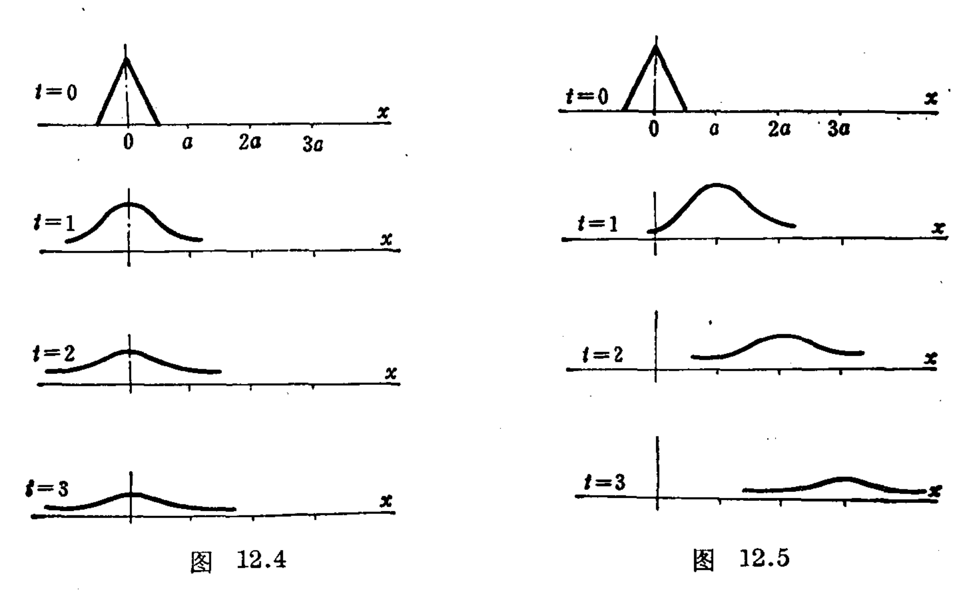


图 12.4　　　　图 12.5

$$T \sim L^2/b \tag{12.1.10}$$

这就是等化时间，也叫做弛豫时间、有生时间、时间常数等等，也就是过程的活跃时间。在实际解扩散问题时，如果没有其它的外部动机，则解题的时段可取为比 T 略大约半个量级即可，例如取 $0 \leqslant t \leqslant 5T$。

对流-扩散方程(12.1.3)在初始条件(12.1.4)下的解为

$$u(x, t) = \frac{1}{\sqrt{4\pi bt}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\xi-at)^2}{4bt}\right] \varphi(\xi) d\xi, \quad -\infty < x < \infty, \quad t \geqslant 0 \tag{12.1.11}$$

图 12.5 表示初始为三角形分布和以后时刻的演化。这里弛豫与波动过程相耦合，因此兼备了两者的特点，一方面有平滑化，另一方面扰动仿佛以集体速度 a 移动，但不能保持波阵面。例如在流体的热交换中，当对流与传导效应并存时就表为方程(12.1.3)，$a\dfrac{\partial u}{\partial x}$ 为对流项，$-b\dfrac{\partial^2 u}{\partial x^2}$ 为扩散项。

§12.2　过程的稳定性和定解条件的恰当性

求解动态过程问题是以初始条件作为主要的定解条件的。由于初始值是由观测或者推算得来的物理量，因而不可避免地含有误差即扰动。如果随着过程的发展扰动会无限增长，则一般说来，要求解这样的问题在物理上就没有意义。因此，初值问题是恰当的即合理的，方程的解必须稳定地依赖于初值，也就是说任何初始扰动在过程发展中将自动衰减或者受控在一定限度之内。对于初始扰动是否稳定是过程的一个重要的内在特征。

设 $u(x, t)$ 是对流方程

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = q \tag{12.2.1}$$

的一个解，对应于初值 $u(x, 0)$。当初值受扰即 $u(x, 0) \sim u(x, 0) + \varepsilon(x, 0)$ 时，相应解也受扰 $u(x, t) \sim u(x, t) + \varepsilon(x, t)$，后者满足方程

$$\frac{\partial(u+\varepsilon)}{\partial t} + a\frac{\partial(u+\varepsilon)}{\partial t} = q$$

两式相减得线性齐次方程

$$\frac{\partial \varepsilon}{\partial t} + a\frac{\partial \varepsilon}{\partial x} = 0 \tag{12.2.2}$$

叫做(12.2.1)的小扰动方程。它的形式与(12.2.1)同，只是右项为 0 成为齐次的。由于初始扰动总可以表成不同频率的正弦波的线性迭加，故可假设初值表为一个正弦波

$$\varepsilon(x, 0) = e^{ikx} \tag{12.2.3}$$

k 为任定实数，为频率参数，通常叫做“波数”，$k/2\pi$ 就是“空间频率”，$\dfrac{1}{k}$ 就是“波长”。试定形如

$$\varepsilon(x, t) = e^{\mu t} e^{ikx} \tag{12.2.4}$$

的解。以(12.2.4)代入(12.2.2)得到 $\mu = \mu(k)$ 的“特征方程”

$$\mu + iak = 0$$

即

$$\mu = -iak \tag{12.2.5}$$

因此得到方程(12.2.2)在初始条件(12.2.3)下的解即谐波解

$$\varepsilon(x,\ t)=e^{-iakt}e^{ikx}=e^{ik(x-at)}$$

由于 a 为实数，不论波长如何，谐波扰动均以速度 a 推进而幅度不变，因此是稳定的。

类似地，对于扩散方程

$$\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}=q \tag{12.2.6}$$

则有小扰动方程组及其谐波解

$$\frac{\partial \varepsilon}{\partial t}-b\frac{\partial^2 \varepsilon}{\partial t^2}=0$$

$$\mu=-bk^2 \qquad \varepsilon(x,\ t)=e^{-bk^2t}e^{ikx} \tag{12.2.7}$$

由于 $b>0$，波形不动而幅度随时间 t 作指数状衰减，波长愈短(即频率 k 愈大)，衰减愈甚。因此也是稳定的。

至于对流-扩散方程，则有

$$\left.\begin{aligned}\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}-b\frac{\partial^2 u}{\partial x^2}=q\\ \frac{\partial \varepsilon}{\partial t}+a\frac{\partial \varepsilon}{\partial x}-b\frac{\partial^2 \varepsilon}{\partial x^2}=0\end{aligned}\right\} \tag{12.2.8}$$

$$\mu=-iak-bk^2 \qquad \varepsilon(x,\ t)=e^{-bk^2t}e^{ik(x-at)} \tag{12.2.9}$$

谐波扰动以速度 a 推进而幅度作指数状衰减。

总结以上各种情况，都有

$$\mathrm{Re}\mu(k)\leqslant 0,\ \ 对一切频率\ k \tag{12.2.10}$$

这是用来判断过程对于初始扰动是否稳定的条件在对流方程，(12.2.5)固然满足(12.2.10)，但特别有

$$\mathrm{Re}\mu(k)=\mathrm{Re}(-ik)=0 \tag{12.2.11}$$

因此可以说是临界稳定或中立稳定，这也是波动过程的共同特点。在扩散方程或扩散-对流方程则有

$$\mathrm{Re}\mu(k)=-bk^2<0 \tag{12.2.12}$$

这也是一切扩散过程的共同特点，表示耗散。

顺便指出，当方程(12.2.6)中的系数 b 为纯虚数时，则成为量子力学中的薛丁格方程，它仍是稳定的(中立稳定)；当 b 为负实数时就是所谓反扩散或反热传导方程，则是不稳定的。当方程(12.2.1)中的 a 为纯虚数时也不稳定，这时方程是椭圆型的。本章中将只讨论稳定初值问题的数值解法。在若干情况下，不稳定初值问题必有特定的物理意义，并且在实践上要求定解。这类问题的数值解法要困难一些，可以参考[2]。

对于初值问题的定解，实践上总是局限在有界的空间区域上，因此还要规定边界条件。关于边界条件也有给得恰当与否的问题。先讨论对流方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0$$

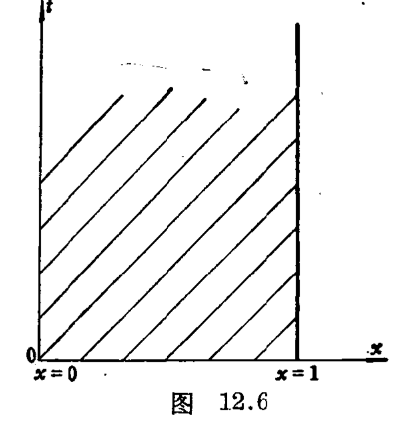


图 12.6

我们知道它的解可以表为 $u(x,\ t)=\varphi(x-at)$，沿特征线 $x-at=\mathrm{const}$ 取常值。假定 $a>0$，在 $t=0$ 上给了初值，要求在 $0\leqslant x\leqslant 1$，$t\geqslant 0$ 上定解(图12.6)。设在左边界即 $x=0$ 上给了 u 的

边值，则由特征线的走向可知在 $0\leqslant x\leqslant 1$，$t\geqslant 0$ 上的解就唯一确定。因此在右边界即 $x=1$ 上就无需也不应该再给边界条件，否则就是过定的，即条件有矛盾或多余。反之，在左边界 $x=0$ 上的边界条件则是不可少的，否则就是欠定的。当 $a<0$ 时情况恰相反，必须在右边界给一个条件而在左边界不给。

对于扩散方程

$$\frac{\partial u}{\partial t}-b\,\frac{\partial^2 u}{\partial x^2}=0$$

微分算子的空间部分为椭圆算子，含有 x 的二阶微商，故应在左右两个边界上各给一个边界条件。通常有给定 u(第一类)，给定 $\frac{\partial u}{\partial x}$(第二类)，给定 u 及 $\frac{\partial u}{\partial x}$ 的组合 (第三类) 三种情况。反之，如果在一边给两个条件，另一边不给条件，则必导致解相对于边值扰动的不稳定性，这样的边界条件是不恰当的。

§12.3 差 分 格 式

偏微分方程的主要数值解法是差分方法。在空间和时间两个方向上将问题离散化为差分方程，然后从初始条件出发，按时间逐层推进。这种方法有高度的通用性，而它的公式又是程式化了的，便于程序实现，因此也叫做差分格式。本节将以对流和扩散方程为例说明有关的一些概念。

将 x-t 平面的上半部 $t\geqslant 0$ 用坐标线

$$x=x_j=j^h,\ j=0,\ \pm 1,\ \pm 2,\ \cdots$$
$$t=t_n=n\tau,\ n=0,\ 1,\ 2,\ \cdots$$

分为格网，h, τ 分别是空间及时间步长。暂时不考虑边界条件的处理及其影响。

对于扩散方程

$$\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}-\varepsilon=0 \tag{12.3.1}$$

用适当的差商代替微商，可以自然地构成种种差分格式。采用记号

$$u_j^n=u(x_j,\ t_n)$$
$$\delta^2 u_j^n=u_{j-1}^n-2u_j^n+u_{j+1}^n$$

并用 E 表示截断误差(其含义见后文)。于是有

$$\left.\begin{aligned}&\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}\,\delta^2 u_j^n-q_j^n=0\\&E=O(\tau)+O(h^2),\quad \text{显式}\end{aligned}\right\} \tag{12.3.2}$$

$$\left.\begin{aligned}&\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}\,\delta^2 u_j^{n+1}-q_j^n=0\\&E=O(\tau)+O(h^2),\quad \text{全隐式}\end{aligned}\right\} \tag{12.3.3}$$

$$\left.\begin{aligned}&\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}(\delta^2 u_j^{n+1}+\delta^2 u_j^n)-q_j^n=0\\&E=O(\tau^2)+O(h^2),\quad \text{平均隐式}\end{aligned}\right\} \tag{12.3.4}$$

$$\left.\begin{aligned}&\frac{1}{2\tau}(u_j^{n+2}-u_j^n)-\frac{b}{h^2}\,\delta^2 u_j^{n+1}-q_j^n=0\\&E=O(\tau^2)+O(h^2),\quad \text{中心差格式}\end{aligned}\right\} \tag{12.3.5}$$

类似地，对于对流方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}-q=0 \tag{12.3.6}$$

则有

$$\left.\begin{array}{c}\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-q_j^n=0\\ E=O(\tau)+O(h^2),\ \text{显式-中心差}\end{array}\right\} \tag{12.3.7}$$

$$\left.\begin{array}{c}\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})-q_j^n=0\\ E=O(\tau)+O(h^2),\ \text{全隐式-中心差}\end{array}\right\} \tag{12.3.8}$$

$$\left.\begin{array}{c}\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_{j+1}^n-u_j^n)-q_j^n=0\\ E=O(\tau)+O(h),\ \text{显式-右偏}\end{array}\right\} \tag{12.3.9}$$

$$\left.\begin{array}{c}\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_j^n-u_{j-1}^n)-q_j^n=0\\ E=O(\tau)+O(h),\ \text{显式-左偏}\end{array}\right\} \tag{12.3.10}$$

当由第 n 个时间层推进到 $n+1$ 层时，公式(12.3.2)中提供了逐点 u_j^{n+1} 的明显表达，因此称为显式。属于显式的还有(12.3.5)，(12.3.7)，(12.3.9～10)。在公式(12.3.3)则需联解一个代数方程组才能得到待定的 $n+1$ 层各点的值，因此叫做隐式，属于隐式的还有(12.3.4)，(12.3.8)。从解算的方便性和工作量来看，显式是有利的。但在某些场合如抛物型方程采用隐式反而更有利(见 §12.6)。

在公式(12.3.2)中，在推算 $n+1$ 层时只用第 n 层的数据，前后联系到二个层次，这叫做双层格式，在程序实现时一般只需保留一片场(一个时层的数据)，其它除(12.3.5)外也都是双层。在(12.3.5)中，是从 n，$n+1$ 两层推算 $n+2$ 层，前后联系到三个层次，叫做三层格式，这是多层格式的一种。在实现时，多层格式需要保留两片场，而且需要另外的方式启步，即从第 0 层推算第 1 层，在这以后才能按该格式进行。在偏微分方程的情况，一般存储量的负担很重，故很少采用超过三层的格式。

将公式(12.3.2)左端各项在 (x_j, t_n) 点作幂次展开，不难验证

$$E=\left\{\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h}\delta^2u_j^n-q_j^n\right\}-\left\{\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}-s\right\}_{x_j,t_n}=O(\tau)+O(h^2)$$

这就是截断误差。因此差分方程的解并不严格满足原来的微分方程而只是近似地满足。但是，正如这个例子表明的，当步长 h，$\tau\to 0$ 时截差 $E\to 0$ 差分方程(12.3.2)的极限形式就是微分方程(12.3.1)，这时我们说差分方程与微分方程相容。这种相容性表示差分方程“收敛”于微分方程。是差分方程的必备条件。通常所谓收敛性，是指差分方程的解，当步长 h，$\tau\to 0$ 时收敛于微分方程的解。相容性与收敛性是不同的概念，前者只是必备的条件，而后者才是最终的目标，在理论分析上要困难些，将不加论述。在许多情况下，差分的相容性再加上稳定性(见 §12.4)可以保证收敛性。

上面列出的差分格式都附有截差式，但幂次展开的基点各有不同。例如(12.3.4)就是在 $(x_j, t_{n+\frac{1}{2}})$ 展开而得的。当 $E=O(\tau^p)+O(h^q)$ 时我们说格式对时间 (τ) 为 p 阶精度，对空间 (h) 为 q 阶精度。

§12.4　差分格式的稳定性

初值问题的差分解法是以步进方式工作的。在逐步推进的过程中，误差也逐步积累。这种误差积累是保持有界还是恶性发展？这就是所谓数值稳定性的问题。数值稳定性是差分格式的必备条件。在不稳定的情况下，寄生误差不仅要湮没真解，而且会导致计算的垮台（如上溢）。因此，一定不稳定的差分格式，即使有其它方面的优点也是不能据以工作的。

先举扩散方程显式格式(12.3.2)为例

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}(u_{j-1}^n-2u_j^n+u_{j+1}^n)-q_j^n=0 \tag{12.4.1}$$

为了简化分析，暂不考虑边界条件的效应，认为 $j=0,\ \pm1,\ \pm2,\ \cdots$展至 $\pm\infty$。设想初值 u_j^0 受扰，即含有误差而成为 $(u+\varepsilon)_j^0=u_j^0+\varepsilon_j^0$，则相应解 u_j^n 也受扰而成为 $(u+\varepsilon)_j^n=u_j^n+\varepsilon_j^n$，它满足与(12.4.1)一样的方程即

$$\frac{1}{\tau}[(u+\varepsilon)_j^{n+1}-(u+\varepsilon)_j^n]-\frac{b}{h^2}[(u+\varepsilon)_{j-1}^n-2(u+\varepsilon)_j^n+(u+\varepsilon)_{j+1}^n]-q_j^n=0$$

与(12.4.1)相减，得到扰动即误差 ε_j^n 所满足的方程

$$\frac{1}{\tau}(\varepsilon_j^{n+1}-\varepsilon_j^n)-\frac{b}{h^2}(\varepsilon_{j-1}^n-2\varepsilon_j^n+\varepsilon_{j+1}^n)=0 \tag{12.4.2}$$

它与(12.4.1)相似，只是除去了 q_j^n 项而成为齐次的线性常系数差分方程。对(12.4.2)可以用谐波分析的方法来定解(与§12.2相仿)。

把初始误差 ε_j^0 表为一个简谐波的形式

$$\varepsilon_j^0=(e^{ikx})_{x=jh}=e^{ikjh}$$

这里 k 为频率参数，即波数，见(12.2.3)$\left(\frac{k}{2\pi}\text{ 就是“空间频率”，}\frac{1}{k}\text{ 就是波长}\right)$试定形如

$$\varepsilon(jh,\ n\tau)=\varepsilon_j^n=(\lambda(k))^n e^{ikjh},\quad j=0,\ \pm1,\ \pm2,\ \cdots;\ n=0,\ 1,\ \cdots \tag{12.4.3}$$

的谐波解。这里 $\lambda=\lambda(k)$ 为对应于波数 k 的增长因子，比较(12.2.4)和(12.4.3)，$\lambda\sim e^{\mu\tau}$，λ 可以用下法定出。将(12.4.3)代入(12.4.2)得

$$\frac{1}{\tau}(\lambda^{n+1}e^{ikjh}-\lambda^n e^{ikjh})=\frac{b}{h^2}(\lambda^n e^{ik(j-1)h}-2\lambda^n e^{ikjh}+\lambda^n e^{ik(j+1)h})=0$$

消去公因子 $\lambda^n e^{ikjh}$ 得方程

$$\frac{1}{\tau}(\lambda-1)-\frac{b}{h^2}(e^{-ikh}-2+e^{ikh})=0 \tag{12.4.4}$$

叫做差分格式(12.4.1)的特征方程。它的根，即特征根，即增长因子

$$\lambda=\lambda(k)=1-\frac{\tau b}{h^2}2(1-\cos kh)=1-4\beta\sin^2\theta \tag{12.4.5}$$

在这里和以后命

$$\beta=\tau b/h^2,\quad \theta=kh \tag{12.4.6}$$

回到(12.4.3)，当 $|\lambda(k)|>1$ 时，误差随 n 作指数状增长，$|\lambda(k)|\leqslant1$ 时则误差不增长。由于初始误差可以表为不同频率 k 的谐波的迭加，并且由于计算中舍入误差的随机性，应该认为所有的 k 的频率组分都是可能出现的。因此数值稳定的条件是

$$|\lambda(k)|\leqslant1,\quad \text{对一切实数 } k \tag{12.4.7}$$

对于(12.4.5)说来，当 k 任意变化时，$4\beta\sin^2\theta$ 变化的范围是$[0,\ 4\beta]$，因此使(12.4.7)成立

的充要条件是

$$\beta\leqslant\frac{1}{2}\quad 即\quad \tau\leqslant h^2/2b \tag{12.4.8}$$

我们说差分格式(12.4.1)是条件稳定的，即当步长 τ, h 满足上列不等式时为稳定，否则不稳定。

不难看出，为了从线性常系数差分方程形成判稳用的特征方程，只需将差分方程中的非齐次项略去，并将 u_{j+s}^{n+p} 项代以 $\lambda^p e^{iksh}$ 即可。据此，可以得到隐式(12.3.3)的特征方程和特征根

$$\frac{1}{\tau}(\lambda-1)-\frac{b}{h^2}(\lambda e^{-ikh}-2\lambda+\lambda e^{ikh})=0,\ \ \lambda=1/(1+4\beta\sin^2\theta)$$

不论 k, τ, h 如何，恒有 $|\lambda|\leqslant1$，这是恒稳的，即无条件稳定。对于(12.3.4)类似地有

$$\lambda=(1-2\beta\sin^2\theta)/(1+2\beta\sin^2\theta)$$

也是恒稳。至于三层中心差格式(12.3.5)，特征方程为二次，有两个特征根

$$\frac{1}{2\tau}(\lambda^2-1)-\frac{b}{h^2}(\lambda e^{-ikh}-2\lambda+\lambda e^{ikh})=0$$

$$\lambda_{1,2}=-4\beta\sin^2\theta\pm\sqrt{1+(4\beta\sin^2\theta)^2}$$

不论 τ, h 如何，总有 k 能使 $|\lambda(k)|>1$，因此是恒不稳定的。尽管这个格式的构成也很“自然”，而且有较高的精度，却是不能工作的。

将上述方法用于对流方程时，特征根即增长因子将得复数。为了方便，命

$$\alpha=a\tau/h,\ \ \theta=kh \tag{12.4.9}$$

对于显式中心差格式(12.3.7)得到

$$\frac{1}{\tau}(\lambda-1)+\frac{a}{2h}(e^{ikh}-e^{-ikh})=0$$

$$\lambda(k)=1-i\alpha\sin\theta,\quad |\lambda(k)|^2=1+\alpha^2\sin^2\theta>1$$

这是对流方程最“自然”的差分格式，却是恒不稳的。至于隐式中心差格式(12.3.8)则有

$$\lambda(k)=1/(1-i\alpha\sin\theta),\quad |\lambda(k)|^2=1/(1+\alpha^2\sin^2\theta)\leqslant1$$

恒稳。

值得注意的是两种偏心差显式。在右偏(12.3.9)时

$$\lambda(k)=1+\alpha-\alpha e^{i\theta}$$

当 k 即 $kh=2\theta$ 变化时，根 $\lambda(k)$ 的轨迹在复数平面上是以 $1+\alpha$ 为中心，半径为 α 的圆。当 $a>0$(即 $\alpha>0$)时，此圆在单位圆之外，因此恒不稳。当 $a<0$ 时，根轨迹圆在单位圆之内的充要条件为

$$|\alpha|\leqslant1\quad 即\quad \tau\leqslant h/|a| \tag{12.4.10}$$

故为条件稳定。至于左偏格式(12.3.10)则恰相反，$a<0$ 时恒不稳，$a>0$ 时条件稳，条件同(12.4.10)。总结起来，在 $a\geqslant0$ 时应取左偏，$a\leqslant0$ 应取右偏，统称为特征型差分格式，因为“差分三角形”的斜边与特征线 $x-at=\text{const}$ 的倾向一致。

微分方程(12.3.1)中的系数 a 表示单向波即扰动的传播速度，$a<0$ 表示自左至右，$a<0$ 表示自右至左。从差分方程也可以引进扰动传播速度的概念。例如在左偏格式(12.3.10)中是从 u_{j-1}^n, u_j^n 计算 u_j^{n+1}，可以认为扰动以速度 $c=h/\tau$ 传播，自左至右。在右偏格式(12.3.9)中则扰动以相同速度传播，但方向相反。上面看到，为了保证稳定性，$a\geqslant0$ 取左

偏，$a\leqslant 0$ 时取右偏，这就意味着差分扰动的传播方向应取得与微分扰动的传播方向相同；而步长条件 (12.4.10) 则意味着差分扰动的传播速度不得落后于微分扰动的传播速度即

$$|c|\leqslant|a| \tag{12.4.11}$$

这一条件通常叫做影响条件或柯朗(Courant)条件，它在双曲型方程差分解法中占有重要的地位，是稳定性的必要条件，在许多情况(但非所有情况)下，它也是稳定性的充分条件。影响条件的意义，通过下面的分析，还可以看得更清楚。

单向波方程(12.1.1)(取 $a>0$)当初值为

$$u(x,\ 0)=\begin{cases}1,\ x<0\\0,\ x\geqslant 0\end{cases}$$

时的解 $u(x,\ t)$ 在直线 $t=0$ 与 $x-at=0$ 之间恒为 0，而在它处恒为 1。设有某个差分格式的扰动传播速度为c。如果影响条件不被满足，例如方向不对头，或者速度落后于 a 即 $-\infty<c<a$。于是差分解 $U(x,\ t)$ 在直线 $t=0$ 与 $x-ct=0$ 之间恒为 0。这样，在直线 $x-ct=0$ 与 $x-at=0$ 的夹角内（见图 12.7）$U\equiv 0$ 而 $U\equiv 1$；即使令 $\tau,\ h\to 0$，只要 $c=h/\tau$ 值保持不变，这种情况就不会改变。因此当影响条件不满足时，差分解就不收敛于微分方程的解。

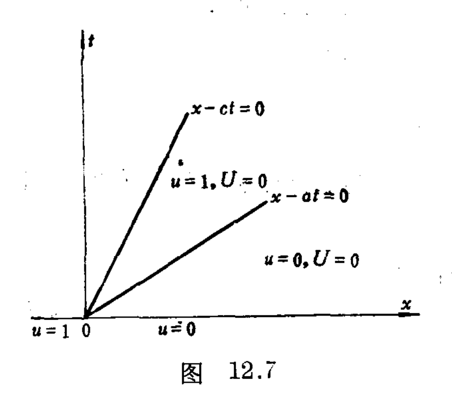


图 12.7

应该指出，差分格式的数值稳定性与微分方程本身的稳定性是有联系但又不同。在 § 12.2 中知道对流方程、扩散方程都是稳定的，但是它们的差分格式却有的稳定，有的不稳定或在一定步长条件下才稳定。要点在于对稳定的微分方程构造稳定的差分格式。

本节为了判稳，采用了谐波分析方法。它在原则上虽只适用于线性常系数方程，但可以适当推广到变系数以及非线性。对于变系数可以采用所谓“冻结”的原则，即对于变系数的一切值都应用谐波分析来判稳得到稳定条件的界限。对于非线性方程，则先将差分格式线性化，然后根据冻结原则用谐波分析法判稳。此外，以上的分析没有考虑到边界条件。边界条件的处理有时不影响稳定性的基本结论，有时则有不良影响。此外，本章中判稳的基本条件取为(12.4.7)，严格说来它只是数值稳定性的必要条件，在许多情况下它还不充分。但是，尽管有上述种种局限性，从这里的初等方法所达到的结论与计算实际是基本相符或接近的，因此具有一定的实用意义。在计算实践上，数值不稳定性大都表现为误差的恶性增长，是很容易察觉的，因此对于复杂问题，也可以用实验测试的手段来决定保证稳定性的步长。关于稳定性问题比较系统的讨论可以参考[1]。

§ 12.5 守恒型差分格式

我们知道，实践上求解的微分方程总是反映物理上的某种守恒律，如质量守恒、动量守恒、能量守恒、粒子数守恒等等。我们以热传导即“温度扩散”为例说明怎样在热量守恒律的基础上导出扩散方程以及怎样在同一基础上导出相应的差分格式。至于双曲方程的情况则见 § 12.8.9。

12.5.1 守恒律的积分形式与微分形式

虽然本章主要讲空间一维或至多为二维的问题，但是为了说明物理背景，以考虑三维空间为便。

设在空间域 Ω 内有温度分布 $u(x, y, z, t)$。分布的不均匀性导致热流，任取一个定向面积元 $d\sigma$，命 $\boldsymbol{n}$ 表示其正向法线。根据傅立叶热传导定律，单位时间内正向通过 $d\sigma$ 的热量为 $-\beta\frac{\partial u}{\partial n}d\sigma$，负号表示热量总是从"热"处流向"冷"处，$\beta>0$ 为介质的热传导系数，可以是常数（均质），可以是变数 $\beta=\beta(x, y, z)$（非均质），甚至可以有间断（组合介质）。

通常情况下，介质单位体积所含的热能正比于温度 u，即 cu；c 为介质按单位体积计算的比热，也和 β 一样，可以是常数或变数甚至有间断。于是体元 dv 内所含热能为 $cudv$。任取子域 $D\subset\Omega$。当时刻 t 由 t' 增至 t'' 时 D 内所有热能的增量为

$$\iiint_D (cu)_{t''}dv-\iiint_D (cu)_{t'}dv$$

图 12.8

这个增量是由下列两种因素引起的：

(1) 在时段 $t'\leqslant t\leqslant t''$ 内通过 D 的边界 ∂D 流入了热量，按照傅立叶传导律，这就是

$$\int_{t'}^{t''}dt\oiint_{\partial D}\beta\frac{\partial u}{\partial n}d\sigma$$

$\boldsymbol{n}$ 表示 ∂D 上外法线方向（图 12.8）。

(2) 在时段 $t'\leqslant t\leqslant t''$ 内区域 D 的内部热源（如果有的话）释放了热量，即

$$\int_{t'}^{t''}dt\iiint_D q dv$$

q 为热源项，表示单位时间单位体积内释放的热量。

于是，根据热量的守恒性应有

$$\iiint_D (cu)_{t''}dv-\iiint_D (cu)_{t'}\cdot dv=\int_{t'}^{t''}dt\oiint_{\partial D}\beta\frac{\partial u}{\partial \nu}d\sigma+\int_{t'}^{t''}dt\iiint_D q dv \tag{12.5.1}$$

对一切 $[t', t'']$，$D\subset\Omega$ 成立。这就是热量守恒律的积分形式，也是最基本的形式。

当有关场量有适当的光滑性时，可以运用高斯积分公式

$$\oiint_{\partial D}\beta\frac{\partial u}{\partial n}d\sigma=\iiint_D\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+\frac{\partial}{\partial z}\beta\frac{\partial u}{\partial z}\right)dv \tag{12.5.2}$$

以及

$$\iiint_D[(cu)_{t''}-(cu)_{t'}]dv=\int_{t'}^{t''}dt\iiint_D\frac{\partial(cu)}{\partial t}dv \tag{12.5.3}$$

因此对一切 $[t', t'']$ 及 $D\subset\Omega$ 有

$$\int_{t'}^{t''}dt\iiint_D\left[\frac{\partial(cu)}{\partial t}-\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+\frac{\partial}{\partial z}\beta\frac{\partial u}{\partial z}+q\right)\right]dv=0 \tag{12.5.4}$$

令 $t', t''\to t$，D 缩到一个点 $(x, y, z)\in\Omega$，就可以"脱括弧"而得扩散方程

$$\Omega:\ \frac{\partial}{\partial t}cu-\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+\frac{\partial}{\partial z}\beta\frac{\partial u}{\partial z}+q\right)=0 \qquad (12.5.5)$$

这是守恒律的微分形式，对于 Ω 内每一点 $(x,\ y,\ z)$ 成立。

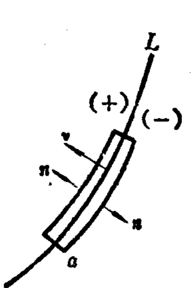


图 12.9

由于高斯积分公式仅当有关场量有适当的光滑性才成立，所以由(12.5.1)至(12.5.5)的过渡并非处处可行。当有介质间断时就是这样。设介质系数 $\beta=\beta(x,\ y,\ z)$ 在曲面 L 上有跳跃性间断，即在其两侧

$$L:\ \beta^{+}\neq\beta^{-}$$

在间断面 L 上守恒律就不表为如(12.5.5)的形式而是采取另外的形式。

取 D 为跨越间断面 L 的任意扁盒状域（其横截面如图 12.9），扁盒的高度 ε 取为小量（相对于"基底"）。对于这样的 D，(12.5.1)当然照样成立。命高度 $\varepsilon\to 0$，扁盒 D 退化为 L 上的一个面元 S，则有

$$\iiint_D (cu)_{t''}dv,\ \iiint_D (cu)_{t'}dv,\ \iiint_D q\,dv\to 0$$

$$\oiint_{\partial D}\beta\frac{\partial u}{\partial n}d\sigma\to\iint_S\left[\left(\beta\frac{\partial u}{\partial\nu}\right)^{+}-\left(\beta\frac{\partial u}{\partial\nu}\right)^{-}\right]d\sigma$$

ν 为 L 上任定的法向，它所指的一方为(+)，另一方为(−)，因此

$$\int_{t'}^{t''}dt'\iint_S\left[\left(\beta\frac{\partial u}{\partial\nu}\right)^{+}-\left(\beta\frac{\partial u}{\partial\nu}\right)^{-}\right]d\sigma=0$$

对于一切 $[t',\ t'']$ 及 $S\subset L$ 成立。命 $t',\ t''\to t$，S 缩到一个点 $(x,\ y,\ z)\in L$ 就可得到

$$L:\ \left(\beta\frac{\partial u}{\partial\nu}\right)^{+}-\left(\beta\frac{\partial u}{\partial\nu}\right)^{-}=0 \qquad (12.5.6)$$

对于 L 上每一点都成立。这就是热量守恒律在介质间断面上的微分形式，通常叫交界条件或间断条件为内边界条件。由于在 L 的两侧 $\beta^{+}\neq\beta^{-}$。故由(12.5.6)得 $\left(\frac{\partial u}{\partial\nu}\right)^{+}\neq\left(\frac{\partial u}{\partial\nu}\right)^{-}$，即法向导数有间断，但其乘积 $\beta\frac{\partial u}{\partial\nu}$ 即热流即通量是连续的。至于扩散方程(12.5.5)本身则在间断面 L 以外之点成立：

$$\Omega-L:\ \frac{\partial}{\partial t}cu-\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+\frac{\partial}{\partial z}\beta\frac{\partial u}{\partial z}+q\right)=0 \qquad (12.5.7)$$



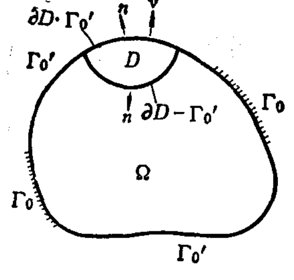


图 12.10

如图 12.10 所示，在 Ω 的边界 $\partial\Omega$ 上，扩散问题的边界条件的一般形式是

$$\partial\Omega=\Gamma_0+\Gamma_0' \qquad (12.5.8)$$

$$\Gamma_0:\ u=\bar{u} \qquad (12.5.9)$$

$$\Gamma_0':\ \beta\frac{\partial u}{\partial\nu}=g(u) \qquad (12.5.10)$$

在 Γ_0 上 u 的值为给定的分布 $\bar{u}$，即第一类边界条件。在 Γ_0' 上流进的通量 $\beta\frac{\partial u}{\partial\nu}$（$\nu$ 为外法向）为给定的函数分布 $g(u)$。例如 $g(u)=p$，即给定通量的值，这就是第二类边界条件，又如 $g(u)$ 为 u 的线性函数 $g(u)=p-\eta u$，$\eta\geqslant 0$，这就是第三类边界条件（包括了第二类），η 为介质与环境之间的热交换系数。

注意第二、三类边界条件，如(12.5.10)，可以自然地吸收在积分守恒律(12.5.1)之中。事实上，任取子域 $D\subset\Omega$，D 的边界 ∂D 可能与 $\partial\Omega$ 上的 Γ_0' 相接触。命 $\partial D\cdot\Gamma_0'$ 表示 ∂D 上属于 Γ_0' 的部分，$\partial D-\Gamma_0'$ 表示 ∂D 上不属于 Γ_0' 的部分，于是 $\partial D=\partial D\cdot\Gamma_0'+(\partial D-\Gamma_0')$

$$\oiint_{\partial D}\beta\frac{\partial u}{\partial n}d\sigma=\iint_{\partial D\cdot\Gamma_0'}\beta\frac{\partial u}{\partial n}d\sigma+\iint_{\partial D-\Gamma_0'}\beta\frac{\partial u}{\partial n}d\sigma=\iint_{\partial D\cdot\Gamma_0'}g(u)d\sigma+\iint_{\partial D-\Gamma_0'}\beta\frac{\partial u}{\partial n}d\sigma$$

因此积分守恒律(12.5.1)就可表为

$$\begin{aligned}&\iiint_D(cu)_{t''}du-\iiint_D(cu)_{t'}du\\&=\int_{t'}^{t''}dt\left[\iint_{\partial D\cdot\Gamma_0'}g(u)d\sigma+\iint_{\partial D-\Gamma_0'}\beta\frac{\partial u}{\partial n}d\sigma+\iiint_D q du\right]\end{aligned}\tag{12.5.11}$$

对一切$[t',t'']$，$D\subset\Omega$ 成立。这时边界条件(12.5.10)已被吸收在内。反过来，在守恒律(12.5.11)的基础上，利用上面采用的“扁盒”方法，可以导出边界条件(12.5.10)。通常称介质间断面的交界条件(如12.5.6)或第二、三类边界条件(如12.5.10)，为自然边界条件，因为它们或者可以自动地从守恒原理导出或者自然地吸收在守恒原理之中。与此相反，第一类边界条件(如12.5.9)则称为强加边界条件。边界条件的这种区分也可以从变分原理的角度来论证，见第十三章。

如果采用向量分析的记号，

$$\operatorname{grad}u=\left(\frac{\partial u}{\partial x},\ \frac{\partial u}{\partial y},\ \frac{\partial u}{\partial z}\right)$$

表示由标量场 u 产生的梯度(向量)场，

$$\operatorname{div}\boldsymbol{A}=\frac{\partial A_x}{\partial x}+\frac{\partial A_y}{\partial y}+\frac{\partial A_z}{\partial z}$$

表示由向量场 $\boldsymbol{A}=(A_x,\ A_y,\ A_z)$ 产生的散度(标量)场，则(12.5.5)可以写成

$$\frac{\partial}{\partial t}cu-\operatorname{div}\beta\operatorname{grad}u-q=0$$

有时也可把(12.5.5)进一步化为

$$\frac{\partial}{\partial t}cu-\left(\beta\Delta u+\frac{\partial\beta}{\partial x}\frac{\partial u}{\partial x}+\frac{\partial\beta}{\partial y}\frac{\partial u}{\partial y}+\frac{\partial\beta}{\partial z}\frac{\partial u}{\partial z}+q\right)=0\tag{12.5.12}$$

$$\Delta=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$$

为了区别于(12.5.12)，可以称方程(12.5.5)的形状为守恒型的或散度型的，因为这是从积分守恒律(12.5.1)过来的原始形式，采取散度的形状。以后将要看到，从数值解的观点，进一步约化为如(12.5.12)的形式是不必要的，最好从原始的守恒型方程(12.5.5)出发，特别是从最为原始的积分守恒律(12.5.1)出发。

在二维的情况下，即一切量不依赖于坐标 z 时，(12.5.1)、(12.5.6)、(12.5.7)分别简化为

$$\iint_D(cu)_{t''}dx\,dy-\iint_D(cu)_{t'}dx\,dy=\int_{t'}^{t''}dt\oint_{\partial D}\beta\frac{\partial u}{\partial n}dS+\int_{t'}^{t''}dt\iint_D q\,dx\,dy\tag{12.5.13}$$

对一切$[t',t'']$，$D\subset\Omega$。

$$\Omega-L:\ \frac{\partial}{\partial t}cu-\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+q\right)=0 \tag{12.5.14}$$

$$L:\ \left(\beta\frac{\partial u}{\partial \nu}\right)^{+}-\left(\beta\frac{\partial u}{\partial \nu}\right)^{-}=0 \tag{12.5.15}$$

这里 Ω 为平面域，L 为介质间断线，在一维情况则进而简化为

$$\int_{x'}^{x''}(cu)_{t''}dx-\int_{x'}^{x''}(cu)_{t'}dx=\int_{t'}^{t''}\left[\left(\beta\frac{\partial u}{\partial x}\right)_{x''}-\left(\beta\frac{\partial u}{\partial x}\right)_{x'}\right]dt+\int_{t'}^{t''}dt\int_{x'}^{x''}q\,dx \tag{12.5.16}$$

对一切 $[t',\ t'']$，$[x',\ x'']\subset\Omega$。

$$\Omega-L:\ \frac{\partial}{\partial t}cu-\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+q\right)=0 \tag{12.5.17}$$

$$L:\ \left(\beta\frac{\partial u}{\partial x}\right)^{+}-\left(\beta\frac{\partial u}{\partial x}\right)^{-}=0 \tag{12.5.18}$$

这 $\Omega=[a,\ b]$ 为线段，L 为 $(a,\ b)$ 内的介质间断点。

12.5.2 守恒律的离散形式

现以一维变系数扩散问题为例来说明怎样构造守恒型的差分格式。在 §12.3 中，是在形式上对于微分方程进行模拟而建立差分方法。但是，归根到底，应该去模拟这个微分方程所反映的守恒律。因此希望差分方程在离散的意义下满足守恒律。

为了这个目的，应该从积分守恒律出发。设在 $x-t$ 平面上求解的区域为 E: $0\leqslant x\leqslant X$，$0\leqslant t\leqslant T$。作格网线(不必等距)：

$$x=x_j,\ j=0,\ 1,\ \cdots,\ J;\quad 0=x_0<x_1<\cdots<x_J=X$$

$$t=t_n,\ n=0,\ 1,\ \cdots,\ N;\quad 0=t_0<t_1<\cdots<t_N=T$$

图 12.11

为了方便引进半线：

$$x=x_{j+\frac{1}{2}}=\frac{1}{2}(x_j+x_{j+1})$$

$$t=t_{n+\frac{1}{2}}=\frac{1}{2}(t_n+t_{n+1})$$

并约定 $x_{-\frac{1}{2}}=x_0$，$x_{J+\frac{1}{2}}=x_J$，于是半线 $x=x_{j+\frac{1}{2}}$ 和整线 $t=t_n$ 把 E 剖分为无遗漏，无重复，无多余的单元(图 12.11)。

$$E_{j}^{n+\frac{1}{2}}=\{x_{j-\frac{1}{2}}\leqslant x\leqslant x_{j+\frac{1}{2}},\ t_n\leqslant t\leqslant t_{n+1}\},\ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N \tag{12.5.19}$$

命

$$C_j^n=\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}(cu)_{t=t_n}dx,\ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N$$

$$B_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\int_{t_n}^{t_{n+1}}\left(\beta\frac{\partial u}{\partial x}\right)_{x=x_{j+\frac{1}{2}}}dt,\ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N-1$$

$$Q_j^{n+\frac{1}{2}}=\int_{t_n}^{t_{n+1}}dt\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}q\,dx,\ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N-1$$

$B_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ 表示通量即由 E_{j+1} 流至 E_j 的热量，而 $-B_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ 则是由 E_j 流至 E_{j+1} 的热量。$B_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ 是在 $E_j^{n+\frac{1}{2}}$ 与 $E_{j+1}^{n+\frac{1}{2}}$ 的交界面 $x=x_{j+\frac{1}{2}}$ 上计值的。

作离散化的第一步，要求积分守恒律(12.5.16)对于一切 $E_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ 成立，即对于一切

$$x'=x_{j-\frac{1}{2}},\ x''=x_{j+\frac{1}{2}},\ t'=t_n,\ t''=t_{n+1},\ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N-1$$

成立。于是

$$C_j^{n+1}-C_j^n=B_{j+\frac{1}{2}}^{n+\frac{1}{2}}-B_{j-\frac{1}{2}}^{n+\frac{1}{2}}+Q_j^{n+\frac{1}{2}},\ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N-1 \quad (12.5.20)$$

作为第二步，将积分 C, B, Q 离散化，命

$$\tau_{n+\frac{1}{2}}=t_{n+1}-t_n,\ h_j=x_{j+\frac{1}{2}}-x_{j-\frac{1}{2}},\ h_{j+\frac{1}{2}}=x_{j+1}-x_j \quad (12.5.21)$$

可以取，例如

$$\left.\begin{aligned}&C_j^n\sim c_j^n u_j^n h_j\\&B_{j+\frac{1}{2}}^{n+\frac{1}{2}}\sim \beta_{j+\frac{1}{2}}^n h_{j+\frac{1}{2}}^{-1}(u_{j+1}^n-u_j^n)\tau_{n+\frac{1}{2}}\\&Q_j^{n+\frac{1}{2}}\sim q_j^n h_j\,\tau_{n+\frac{1}{2}}\end{aligned}\right\} \quad (12.5.22)$$

代入(12.5.21)得内点的守恒差分格式

$$(C_j^{n+1}u_j^{n+1}-C_j^n u_j^n)h_j=[\beta_{j+\frac{1}{2}}^n h_{j+\frac{1}{2}}^{-1}(u_{j+1}^n-u_j^n)-\beta_{j-\frac{1}{2}}^n h_{j-\frac{1}{2}}^{-1}(u_j^n-u_{j-1}^n)]\tau_{n+\frac{1}{2}}+q_j^n h_j\tau_{n+1},$$

$$j=1,\ \cdots,\ J-1;\ n=0,\ \cdots,\ N-1 \quad (12.5.23)$$

在常系数 $c_j^n\equiv c=1$, $\beta_{j+\frac{1}{2}}^n\equiv bc$ 和等距格网 $h_j\equiv h_{j+\frac{1}{2}}\equiv h$, $\tau_{n+\frac{1}{2}}=\tau$ 时这就是显式格式(12.3.2)。

这样的差分格式即使对于介质间断也是适应的，只须在分格网的注意把介质间断点落在整点 x_j 上，此外无须作任何处理。这是因为，介质间断点上的交界条件(12.5.18)本身也是守恒律(12.5.16)导来的。

当区间[0, X]的两端是第一类边界条件

$$u=\bar{u} \quad (12.5.24)$$

时，则增补两个边界方程

$$u_0^{n+1}=\bar{u}_0^{n+1},\ u_J^{n+1}=\bar{u}_J^{n+1},\ n=0,\ 1,\ \cdots,\ N-1 \quad (12.5.25)$$

当两端是自然边界条件即

$$\beta\frac{\partial u}{\partial\nu}=g(u)\equiv p-\eta u \quad (12.5.26)$$

$\frac{\partial u}{\partial\nu}$ 表示对区间[0, X]的外向微商。这就是说

$$\left.\begin{aligned}&x=0=x_0:\ -\beta\frac{\partial u}{\partial x}=g_0(u_0)=p_0-\eta_0u_0\\&x=X=x_J:\ \beta\frac{\partial u}{\partial x}=g_J(u_J)=p_J-\eta_Ju_J\end{aligned}\right\} \quad (12.5.27)$$

由于自然边界条件可以吸收在守恒律之中(见(12.5.11))，考虑到约定 $x_{-\frac{1}{2}}=x_0$, $x_{J+\frac{1}{2}}=x_J$，故有

$$\left.\begin{aligned}&B_{-\frac{1}{2}}^{n+\frac{1}{2}}=\int_{t_n}^{t_{n+1}}\left(\beta\frac{\partial u}{\partial x}\right)_{x=x_0}dt\approx -g_0(u_0^n)\tau_{n+\frac{1}{2}}\\&B_{J+\frac{1}{2}}^{n+\frac{1}{2}}=\int_{t_n}^{t_{n+1}}\left(\beta\frac{\partial u}{\partial x}\right)_{x=x_J}dx\approx g_J(u_J^n)\tau_{n+\frac{1}{2}}\end{aligned}\right\} \quad (12.5.28)$$

因此与内点一样，对于 $E_0^{n+\frac{1}{2}}$, $E_J^{n+\frac{1}{2}}$ 运用守恒律而得两个边界点的差分格式

$$\left.\begin{aligned}(C_0^{n+1}u_0^{n+1}-C_0^n u_0^n)h_0&=[\beta_{\frac{1}{2}}^n h_{\frac{1}{2}}^{-1}(u_1^n-u_0^n)+g_0(u_0^n)]\tau_{n+\frac{1}{2}}+q_0^n h_0\tau_{n+\frac{1}{2}},\\ h_0&=x_{\frac{1}{2}}-x_0\\ (C_J^{n+1}-u_J^{n+1}-C_J^n u_J^n)h_J&=[g_J(u_J^n)-\beta_{J-\frac{1}{2}}^n h_{J-\frac{1}{2}}^{-1}(u_J^n-u_{J-1}^n)]\tau_{n+\frac{1}{2}}+q_J^n h_J\tau_{n+\frac{1}{2}},\\ h_J&=x_J-x_{J-\frac{1}{2}}\end{aligned}\right\}\qquad(12.5.29)$$

第二、三类边界条件已经吸收在内。

以上的差分格式(12.5.23)，(12.5.29)只是表示守恒律(12.5.16)对每个单元 $E_j^{n+\frac{1}{2}}$ 成立。事实上由此就可以保证这种离散的守恒性在更大的范围内也成立。例如，取 $x'=x_{j'-\frac{1}{2}}$，$x''=x_{j''+\frac{1}{2}}$，$t'=t_{n'}$，$t''=t_{n''}$，考虑区域

$$\{x'\leqslant x\leqslant x'',\ t'\leqslant t\leqslant t''\}=\sum_{j=j'}^{j''}\sum_{n=n'}^{n''}E_j^{n+\frac{1}{2}}\qquad(12.5.30)$$

将方程(12.5.23)即(12.5.20)按 $j=j',\ \cdots,\ j''$；$n=n',\ \cdots,\ n''$ 累加

$$\sum_{j=j'}^{j''}\sum_{n=n'}^{n''}(C_j^{n+1}-C_j^n)=\sum_{j=j'}(B_{j+\frac{1}{2}}^{n+\frac{1}{2}}-B_{j-\frac{1}{2}}^{n+\frac{1}{2}}+Q_j^{n+\frac{1}{2}})$$

注意所有内部交界项均互相抵消，因此得到

$$\sum_{j=j'}^{j''}C_j^{n''}-\sum_{j=j'}^{j''}C_j^{n'}=\sum_{n=n'}^{n''}B_{j''+\frac{1}{2}}^{n+\frac{1}{2}}-\sum_{n=n'}^{n''}B_{j'-\frac{1}{2}}^{n+\frac{1}{2}}+\sum_{j=j'}^{j''}\sum_{n=n'}^{n''}Q_j^{n+\frac{1}{2}}$$

这就相当于区域(12.5.30)上的守恒关系(12.5.16)。这样，从每个单元的守恒性出发，由于相邻单元的交界面项都正负相消——也就是说在差分格式上保证了甲方支付给乙方的总是等于乙方从甲方收到的——结果自动保证了大范围的守恒性。守恒要点正在于此。

差分方程(12.5.23)，(12.5.29)是与扩散方程(12.5.17)，交界条件(12.5.18)以及边界条件(12.5.27)相容的。见§12.3 事实上，设 $x=x_j$ 不是介质间断点，将(12.5.23)除以 $h_j\tau_{n+\frac{1}{2}}$ 得到

$$\tau_{n+\frac{1}{2}}^{-1}(c_j^{n+1}u_j^{n+1}-c_j^n u_j^n)-h_j^{-1}\,[\beta_{j+\frac{1}{2}}h_{j+\frac{1}{2}}^{-1}(u_{j+1}^n-u_j^n)-\beta_{j-\frac{1}{2}}h_{j-\frac{1}{2}}^{-1}(u_j^n-u_{j-1}^n)]-q_j^n=0$$

命 h_j，$h_{j+\frac{1}{2}}$，$h_{j-\frac{1}{2}}$，$\tau_{n+\frac{1}{2}}\to 0$ 则其极限形式就是(12.5.17)。设 $x=x_j$ 是介质间断点，在其两侧 $b_j^+\neq b_j^-$。当 $h_{j-\frac{1}{2}}$，$h_{j-\frac{1}{2}}$，$h_{j+\frac{1}{2}}\to 0$ 则有

$$(C_j^{n+1}u_j^{n+1}-C_j^n u_j^n)h_j\to 0,\ \beta_{j+\frac{1}{2}}h_{j+\frac{1}{2}}^{-1}(u_{j+1}^n-u_j^n)\to\left(\beta\frac{\partial u}{\partial x}\right)_j^+,\ \beta_{j-\frac{1}{2}}h_{j-\frac{1}{2}}^{-1}(u_j^n-u_{j-1}^n)\to\left(\beta\frac{\partial u}{\partial x}\right)_j^-$$

因此(12.5.23)的极限形式就是交界条件(12.5.18)。类似地取方程(12.5.29)，命 h_0，h_J，$h_{\frac{1}{2}}$，$h_{J-\frac{1}{2}}\to 0$，则得边界条件(12.5.27)。突出之点在于，我们直接从积分守恒律出发，并没有利用方程(12.5.17)及(12.5.18)和(12.5.27)而得到与之相容的差分方程组。这当然是理应如此的，因为守恒微分方程连同自然边界条件与这里的差分方程组一样都是积分守恒律的推论，前者是守恒律的微分形式，后者是守恒律的离散形式即差分形式。

守恒差分格式的优点在于它在离散的意义下使得守恒律得到严格满足，对于交界条件以及第二、三类边界条件等所谓自然边界条件是采取"自然"处理的方式，不是形式地，孤立地处理而是与守恒律统一在一起。这些特点在物理上是比较令人满意的，当问题复杂时，这种优点更显著，此外，这种差分格式在形式上的单一性对于程序实现和解算也是有利的。

隐式的守恒差分格式也可以用类似原则来推导。例如改取

$$B_{j+\frac{1}{2}}^{n+\frac{1}{2}} \sim \beta_{j+\frac{1}{2}}^{n+1} h_{j+\frac{1}{2}} (u_{n+1}^{n+1} - u_j^{n+1}),\ Q_j^{n+\frac{1}{2}} \sim q_j^{n+1} h_j \tau_{n+\frac{1}{2}}$$

或取

$$B_{j+\frac{1}{2}}^{n+\frac{1}{2}} \sim \left[\frac{1}{2}\beta_{j+\frac{1}{2}}^{n+1} h_{j+\frac{1}{2}}^{-1} (u_{j+1}^{n+1} - u_j^{n+1}) + \frac{1}{2}\beta_{j+\frac{1}{2}}^{n} h_{j+\frac{1}{2}}^{-1} (u_{j+1}^{n} - u_j^{n})\right]\tau_{n+\frac{1}{2}}$$

$$Q_j^{n+\frac{1}{2}} \sim \left(\frac{1}{2} q_j^{n+1} + \frac{1}{2} q_j^{n}\right) h_j \tau_{n+\frac{1}{2}}$$

这就相当于全隐式(12.3.3)和平均隐式(12.3.4)。

也可以从守恒型微分方程(12.5.17)出发来形成差分格式，例如取

$$\left(\frac{\partial}{\partial t} cu\right)_j^n \sim \frac{1}{\tau_{n+\frac{1}{2}}} (c_j^{n+1} u_j^{n+1} - c_j^n u_j^n)$$

$$\left(\frac{\partial}{\partial x} \beta \frac{\partial u}{\partial x}\right)_j^n \sim \frac{1}{h_j}\left[\left(\beta \frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}}^n - \left(\beta \frac{\partial u}{\partial x}\right)_{j-\frac{1}{2}}^n\right]$$

$$\left(\beta \frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}}^n \sim \beta_{j+\frac{1}{2}} \frac{1}{h_{j+\frac{1}{2}}} (u_{j+1}^n - u_j^n)$$

$$(q)_j^n = q_j^n$$

则得到与(12.5.23)相同的守恒格式，上面已经见到，当介质间断点落在节点上，这也与交界条件相适应而无须特别处理。上面介绍的基于积分守恒律的方法，看来冗长一些，但其本质是简单的，物理意义也更明确。在比较复杂的情况下，例如二维问题，几何形状复杂，不等距格网介质间断条件复杂等情况下，采用“积分”的方法是更为有利和可靠的。

习惯上也有从方程(12.5.17)的非守恒型(设 c, b 与 t 无关)

$$c\frac{\partial u}{\partial t} - \beta \frac{\partial^2 u}{\partial x^2} - \frac{\partial \beta}{\partial x}\cdot\frac{\partial u}{\partial x} - q = 0 \tag{12.5.31}$$

出发来构造差分格式，例如取(等距格网)

$$c_j \tau^{-1} (u_j^{n+1} - u_j^n) = \beta_j h^{-2} (u_{j-1}^n - 2u_j^n + u_{j+1}^n) - (2h)^{-1} (\beta_{j+1} - \beta_{j-1}) (2h)^{-1} (u_{j+1}^n - u_{j-1}^n) + q_j^n$$

这个格式就不严格守恒。事实上，将它改写为

$$c_j \tau^{-1} (u_j^{n+1} - u_j^n) = h^{-1}\left[\left(\beta_j + \frac{\beta_{j+1} - \beta_{j-1}}{4}\right) h^{-1} (u_{j+1}^n - u_{j-1}^n)\right.$$

$$\left. - \left(\beta_j - \frac{\beta_{j+1} - \beta_j}{4}\right) h^{-1} (u_j^n - u_{j-1}^n)\right] - q_j^n$$

对于节点 x_{j+1} 也可列类似的方程，只须将 j 改为 $j+1$。于是单元 E_{j+1} 给予 E_j 的热量 $\left(\beta_j + \frac{\beta_{j+1} - \beta_{j-1}}{4}\right) h^{-1} (u_{j+1}^n - u_j^n)$ 一般地不等于单元 E_j 受之于 E_{j+1} 的热量 $\left(\beta_{j+1} - \frac{\beta_{j+2} - \beta_j}{4}\right) \times (u_{j+1}^n - u_j^n)$ (当 b_j 不为常数时)。不过，当系数 β 光滑的时候，这种偏差随 $h \to 0$ 而 $\to 0$。但是当系数 b 有间断时就会造成重大的偏差。在介质间断处，方程(12.5.31)根本不成立，必须代之以交界条件 (12.5.18) 而对后者进行离散化，这样做又引起场内各点格式形式的不统一，而且比较繁琐。一般说来，在非守恒型方程的基础上来离散化反而不太方便，而且容易导致差错。

§ 12.6 扩散方程的差分格式

在§12.3中曾对一维扩散方程列出几种差分格式，在此基础上还可派生出另外几种实用的格式，并列为本节之末表12.1。这些格式对于二维的推广以及根据二维特点设计的一些特殊格式则列为表12.2。表中还附列了截断误差E以及增长因子和稳定条件，供参考。我们将不对表中的格式逐一进行分析论证，而只将一些有关问题择要说明如下。

表12.1 一维扩散方程差分格式

$$\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}-q=0$$

$$\delta^2 u_j^n=u_{j-1}^n-2u_j^n+u_{j+1}^n,\quad \beta=\tau b/h^2,\quad \theta=kh$$

名　称	格　式　与　截　差	增长因子与稳定条件
显　式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}\delta^2u_j^n-q_j^n=0$ $E=O(\tau+h^2)$	$\lambda=1-4\beta\sin^2\frac{\theta}{2}$ $\beta\leqslant\frac{1}{2}$
全隐式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}\delta^2u_j^{n+1}-q_j^{n+1}=0$ $E=O(\tau+h^2)$	$\lambda=\left(1-4\beta\sin^2\frac{\theta}{2}\right)^{-1}$ 恒　稳
平　均 隐　式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{1}{2}\left[\frac{b}{h^2}\delta^2u_j^n-q_j^n\right]-\frac{1}{2}\left[\frac{b}{h^2}\delta^2u_j^{n+1}-q_j^{n+1}\right]=0$ $E=O(\tau^2+h^2)$	$\lambda=\left(1-2\beta\sin^2\frac{\theta}{2}\right)\left(1+2\beta\sin^2\frac{\theta}{2}\right)^{-1}$ 恒　稳
加　权 隐　式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\sigma\left[\frac{b}{h^2}\delta^2u_j^n-q_j^n\right]-(1-\sigma)\left[\frac{b}{h^2}\delta^2u_j^{n+1}-q_j^{n+1}\right]=0$	$\lambda=\left(1-4\sigma\sin^2\frac{\theta}{2}\right)\left(1+4(1-\sigma)\sin^2\frac{\theta}{2}\right)^{-1}$ $0\leqslant\sigma\leqslant\frac{1}{2}$ 时恒稳 $\frac{1}{2}<\theta\leqslant1$ 时 $\beta\leqslant\frac{1}{2(2\sigma-1)}$
中心差	$\frac{1}{2\tau}(u_j^{n+2}-u_j^n)-\frac{b}{h^2}\delta u_j^{n+1}-q_j^{n+1}=0$ $E=O(\tau^2+h^2)$	$\lambda^2+(4\beta-4\beta\cos\theta)\lambda+1=0$ $\lambda_{1,2}=-4\beta\sin^2\frac{\theta}{2}\pm\sqrt{1+\left(4\beta\sin^2\frac{\theta}{2}\right)^2}$ 恒　不　稳
菱　形	$\frac{1}{2\tau}(u_j^{n+2}-u_j^n)-\frac{b}{h^2}(u_{j-1}^{n+1}-u_j^{n+2}-u_j^n+u_{j+1}^{n+1})-q_j^{n+1}=0$ $E=O(\tau^2/h^2)+O(\tau^2+h^2)$	$(1+2\beta)\lambda^2-(4\beta\cos\theta)\lambda-(1-2\beta)=0$ $\lambda_{1,2}=(2\beta\cos\theta\pm\sqrt{1-4\beta^2\sin^2\theta})(1+2\beta)$ 恒　稳
跳　点	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}\delta^2u_j^n-q_j^n=0$，当 $n+1+j=$偶 $\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}\delta^2u_j^{n+1}-q_j^{n+1}=0$，当 $n+1+j=$奇 $E=O(\tau^2/h^2)+O(\tau^2+h^2)$	同　上

(1) 显式和隐式的比较

差分格式中步长τ, h的选取要受到截断误差和稳定性两个方面的制约。扩散方程显式格式的稳定条件是$\tau\leqslant h^2/2b$，即$\tau=O(h^2)$，时间步长τ为空间步长h的二阶小量。这是相当苛刻的条件。例如，为了提高精度，把h减半，则τ必须缩小四倍，从而总工作量增加八倍。对于隐式，每一时间步要联解一个代数方程组。在一维情况，系数阵的三对角线带状阵，有象追赶法这样有效的解法，使每步联解的工作量仅线性地依赖于节点数，与显式工作

表 12.2 二维扩散方程差分格式

$$\frac{\partial u}{\partial t}-b\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)-q=0$$

$$\delta_x^2 u_{ij}^n=u_{i-1,j}^n-2u_{ij}^n+u_{i+1,j}^n,\quad \beta_1=\tau b/h_1^2,\quad \theta_1=k_1h_1$$

$$\delta_y^2 u_{ij}^n=u_{i,j-1}^n-2u_{ij}^n+u_{i,j+1}^n,\quad \beta_2=\tau b/h_2^2,\quad \theta_2=k_2h_2$$

名 称	格式与截差	增长因子与稳定条件
显 式	$\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^n)-\frac{b}{h_1^2}\delta_x^2u_{ij}^n-\frac{b}{h_2^2}\delta_y u_{ij}^n-q_{ij}^n=0$ $E=O(\tau+h_1^2+h_2^2)$	$\lambda=1-4\left(\beta_1\sin^2\frac{\theta_1}{2}+\beta_2\sin^2\frac{\theta}{2}\right)$ $\tau\leqslant\left[2b\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right)\right]^{-1}$
全隐式	$\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^n)-\frac{b}{h_1^2}\delta_x^2u_{ij}^{n+1}-\frac{b}{h_2^2}\delta_y^2u_{ij}^{n+1}-q_{ij}^{n+1}=0$ $E=O(\tau+h_1^2+h_2^2)$	$\lambda=\left[1+4\left(\beta_1\sin^2\frac{\theta_1}{2}+\beta_2\sin^2\frac{\theta_2}{2}\right)\right]^{-1}$ 恒 稳
平 均 隐 式	$\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^n)-\frac{1}{2}\left[\frac{b}{h_1^2}\delta_x^2u_{ij}^n+\frac{b}{h_2^2}\delta_y^2u_{ij}^n-q_{ij}^n\right]$ $-\frac{1}{2}\left[\frac{b}{h_1}\delta_x^2u_{ij}^{n+1}+\frac{b}{h_2^2}\delta_y^2u_{ij}^{n+1}-q_{ij}^{n+1}\right]=0$	$\lambda=\left[1-2\left(\beta_1\sin^2\frac{\theta_1}{2}+\beta_2\sin^2\frac{\theta_2}{2}\right)\right]$ $\cdot\left[1+2\left(\beta_1\sin^2\frac{\theta_1}{2}+\beta_2\sin^2\frac{\theta_2}{2}\right)\right]^{-1}$ 恒 稳
菱 形	$\frac{1}{2\tau}(u_{ij}^{n+2}-u_{ij}^n)-\frac{b}{h_1^2}(u_{i-1,j}^{n+1}-u_{ij}^{n+2}-u_{ij}^n+u_{i+1,j}^{n+1})$ $-\frac{b}{h_2^2}(u_{i,j-1}^{n+1}-u_{ij}^{n+2}-u_{ij}^n+u_{i,j+1}^{n+1})-q_{ij}^{n+1}=0$ $E=O\left(\frac{\tau^2}{h_1^2}+\frac{\tau^2}{h_2^2}\right)+O(\tau^2+h_1^2+h_2^2)$	恒 稳
跳 点	$\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^n)-\frac{b}{h_1^2}\delta_x^2u_{ij}^n-\frac{b}{h_2^2}\delta_y^2u_{ij}^n-q_{ij}^n=0,n+1+i+j=$偶 $\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^n)-\frac{b}{h_1^2}\delta_x^2u_{ij}^{n+1}-\frac{b}{h_2^2}\delta_y^2u_{ij}^{n+1}-q_{ij}^{n+1}=0,n+1+i+j=$奇 $E=O\left(\frac{\tau^2}{h_1^2}+\frac{\tau^2}{h_2^2}\right)+O(\tau^2+h_1^2+h_2^2)$	同 上
交 替 方 向	$\frac{1}{\tau}(u_{ij}^{n+\frac{1}{2}}-u_{ij}^n)-\frac{1}{2}\left[\frac{b}{h_1^2}\delta_x^2u_{ij}^{n+1}+\frac{b}{h_2^2}\delta_y^2u_{ij}^n-q_{ij}^n\right]=0$ $\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^{n+\frac{1}{2}})-\frac{1}{2}\left[\frac{b}{h_1^2}\delta_x^2u_{ij}^{n+\frac{1}{2}}+\frac{b}{h_2^2}\delta_y^2u_{ij}^{n+1}-q_{ij}^{n+1}\right]=0$ $E=O(\tau+h_1^2+h_2^2)$	$\lambda_1=\left(1-\beta_1\sin^2\frac{\theta_1}{2}\right)\cdot\left(1+\beta_2\sin^2\frac{\theta_2}{2}\right)^{-1}$ $\lambda_2=\left(1-\beta_2\sin^2\frac{\theta_2}{2}\right)\cdot\left(1+\beta_1\sin^2\frac{\theta_1}{2}\right)^{-1}$ $\lambda=\lambda_1\lambda_2$，恒稳
局 部 一 维	$\frac{1}{\tau}(u_{ij}^{n+\frac{1}{2}}-u_{ij}^n)-\frac{b}{h_1^2}\delta_x^2u_{ij}^{n+\frac{1}{2}}-\frac{1}{2}q_{ij}^n=0$ $\frac{1}{\tau}(u_{ij}^{n+1}-u_{ij}^n)-\frac{b}{h_2^2}\delta_y^2u_{ij}^{n+1}-\frac{1}{2}q_{ij}^n=0$ $E=O(\tau+h_1^2+h_2^2)$	$\lambda_1=\left(1+4\beta_1\sin^2\frac{\theta_1}{2}\right)^{-1}$ $\lambda_2=\left(1+4\beta_2\sin^2\frac{\theta_2}{2}\right)^{-1}$ $\lambda=\lambda_1\lambda_2$，恒稳

量相比量级相同或接近。由于隐式的恒稳性，τ 的选取不受稳定性的限制而只决定了截差。因此有可能通过 τ 的放大而节约工作量。

此外，从扩散过程的物理特征来看也以取隐式为好。扩散的影响是瞬刻传开的在任意时刻 t 任意坐标 x 的状态受到初始 $t=0$ 时全轴 $-\infty<x<\infty$ 的影响的，也就是说扰动以无限大的速度传播。当采用显式时，扰动在差分格网中的传播速度是 h/τ，因此如取 $\tau=O(h)$，则它始终将落后于实际扰动的传播，故不稳定。反之，在采用隐式时，在每个时间步内，每个节点值影响全部节点，即扰动也是瞬刻传开的，这就比较符合于物理模型，同时也保证了稳定性。

从截差的角度来看，平均隐式比全隐式更有利。对全隐式，$E=O(\tau)+O(h^2)$，如取 $\tau=O(h)$ 则 $E=O(h)$ 为一阶精度；为要达到二阶精度则应取 $\tau=O(h^2)$，这和显式的步长条件相当，隐式的好处也就抵消了。反之，在平均隐式，由于恒稳可取 $\tau=O(h)$，于是 $E=O(\tau^2)+O(h^2)=O(h^2)$，这就达到了二阶精度。

在计算实践中解一维扩散方程主要是采用隐式，特别是平均隐式，也叫做 Crank-Nicolson 格式。

(2) 解一维隐式的追赶法

在一维扩散方程的隐式格式，配合着适当的边界处理，在每个时间步要解一个线代数方程，其系数矩阵为三对线带状

$$\begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{N-1} & b_{N-1} & c_{N-1} \\ & & & a_N & b_N \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{bmatrix} \tag{12.6.1}$$

对此运用顺序的高斯消元法即得追赶法。该法分为正消和反代两个过程。

正消：置 $\beta_0=0,\ v_0=0$

对于 $k=1,\ 2,\ \cdots,\ N$

$$\alpha_k=a_k\beta_{k-1}+b_k$$

$$\beta_k=-\alpha_k^{-1}c_k$$

$$v_k=\alpha_k^{-1}(f_k-a_kv_{k-1})$$

反代：置 $u_{N+1}=0$

对于 $k=N,\ N-1,\ \cdots,\ 2,\ 1$

$$u_k=v_k-\beta_ku_{k+1}$$

追赶法中正反两过程都是步进式，也有稳定性的问题。在扩散方程，矩阵的对角元占优势，可以证明追赶过程是稳定的。

(3) 菱形格式与跳点格式

在 §12.4 中见到，扩散方程

$$\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}-q=0 \tag{12.6.2}$$

的中心差显式格式

$$\frac{1}{2\tau}(u_j^{n+2}-u_j^n)-\frac{b}{h^2}(u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1})-q_j^{n+1}=0$$

$$E=O(\tau^2)+O(h^2)$$

是一种很自然的逼近，具有较高的二阶精度，但恒不稳，不能使用。但是，如将上式中的 u_j^{n+1} 换以 $\frac{1}{2}(u_j^{n+2}+u_j^n)$ 就变成一个恒稳的实用格式

$$\frac{1}{2\tau}(u_j^{n+2}-u_j^n)-\frac{b}{h^2}(u_{j-1}^{n+1}-u_j^{n+2}-u_j^n+u_{j+1}^{n+1})-q_j^{n+1}=0 \quad \Diamond \tag{12.6.3}$$

公式中只用到节点 $(j,\ n)$，$(j,\ n+2)$，$(j-1,\ n+1)$，$(j+1,\ n+1)$ 形成菱形，但不涉及中点

$(j,\ n+1)$，是一个“空心”的菱形，因此叫做菱形格式，或称 Dufort-Frankel 格式，这是三层的，特征方程为二次

$$(1+2\beta)\lambda^2-(4\beta\cos\theta)\lambda-(1-2\beta)=0,\ \beta=\tau b/h^2,\ \theta=kh \tag{12.6.4}$$

不难验算有两个特征根，其模量恒 $\leqslant 1$

$$\lambda_{1,2}=(2\beta\cos\theta\pm\sqrt{1-4\beta^2\sin^2\theta})/(1+2\beta),\ |\lambda_{1,2}|\leqslant 1 \tag{12.6.5}$$

因此恒稳，加上它是显式的，这是它的主要优点。由于它是三层的，故也具有三层格式的共同缺点，即要求两片场的存储量和另法启步。但是，它的主要缺点在于相容性和精确度的问题。

事实上，在节点 $(j,\ n+1)$ 作幂次展开，得到截断误差

$$E=\left(\frac{\tau}{h}\right)^2 b\,\frac{\partial^2 u}{\partial t}+O\left(\tau^2+h^2+\frac{\tau^4}{h^2}\right)=O\left(\frac{\tau^2}{h^2}\right)+O\left(\tau^2+h^2+\frac{\tau^4}{h^2}\right) \tag{12.6.6}$$

设想在 $h,\ \tau\to 0$ 时 $\tau/h=r$ 保持常值，则菱形差分方程的极限形式是带阻尼的波动方程

$$r^2 b\,\frac{\partial^2 u}{\partial t^2}+\frac{\partial u}{\partial t}-b\,\frac{\partial^2 u}{\partial x^2}-q=0 \tag{12.6.7}$$

而不是扩散方程。为了保证差分格式对扩散方程的相容性，必需要求当 $h,\ \tau\to 0$ 时 τ/h 也 $\to 0$。为此，可以取为 $\tau=O(h^{1+s}),\ s>0$。

由此可见，在菱形法中，时间步长 τ 虽不受稳定性的限制，但却受相容条件的限制。后者 $\tau=O(h^{1+s})$。在 $1>s>0$ 时虽比显式的稳定条件 $\tau=O(h^2)$ 稍宽，但 $E=O(h^s)+O(\tau^2+h^2)$ 比显式的 $E=O(h)+O(\tau^2)$ 降低了精度。

设想将时空节点 $(j,\ n)$ 按照 $n+j=$ 奇数或偶数分为两组，形成两套互相交错的菱形格网如图 12.12，不难看出，在菱形算法的推进过程中，这两套网点是互相独立的。

图 12.12

所谓跳点法就是在上述奇偶分组的基础上进行的，当从第 n 层进至第 $n+1$ 层时，先在偶点用显式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}(u_{j-1}^n-2u_j^n+u_{j+1}^n)-q_j^n=0,\quad n+1+j=\text{偶} \tag{12.6.8}$$

产生新值。然后在奇点用全隐式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)-\frac{b}{h^2}(u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1})-q_j^{n+1}=0,\quad n+1+j=\text{奇} \tag{12.6.9}$$

这时奇点 u_j^{n+1} 的左右邻 u_{j-1}^{n+1}，u_{j+1}^{n+1} 都是偶点，已经有了新值。因此这是奇、偶、显、隐交替的方法，其中隐式只是形式上的，实质上还是显的。

可以把这个算法稍加变形以节约工作量。事实上，每当用隐式 (12.6.9) 算出奇点值 u_j^{n+1} 时，由于 $n+2+j$ 必为偶数，故由 12.6.8 有显式表达

$$\frac{1}{\tau}(u_j^{n+2}-u_j^{n+1})-\frac{b}{h^2}(u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1})-q_j^{n+1}=0,\ n+2+j=\text{偶} \tag{12.6.10}$$

将此与(12.6.9)相减，得到

$$u_j^{n+2}=2u_j^{n+1}-u_j^n,\ n+2+j=\text{偶} \tag{12.6.11}$$

故奇点值 u_j^{n+1} 算出后无须保存而直接用只含两个加法的公式(12.6.11)代替 (12.6.8) 以产生下一层的偶点值 u_j^{n+2}，后者则被保存。因此，在初始层 u_j^0 的基础上，首先对层 $n=1$ 偶点

值用显式(12.6.8)产生，以后便按方才所说的方式进行，直至算出最后第 $n=N$ 层的偶点值，最后用隐式(12.6.9)补出该层的奇点值。

注意，从式(12.6.9)及式(12.6.11)消去 u_j^{n+1}，得到

$$\frac{1}{2\tau}(u_j^{n+2}-u_j^n)-\frac{b}{h^2}(u_{j-1}^{n+1}-u_j^{n+2}-u_j^n+u_{j+1}^{n+1})-q_j^{n+1}=0,\ n+j=\text{偶}$$

这就是菱形公式(12.6.3)。但它只在偶点格网进行而弃去了奇点格网，在菱形法中这奇偶两套本来是互相独立的。因此，跳点法实质上与菱形法等价，具有相同的稳定性(12.6.4～5)和截断误差(12.6.6)，但在算法组织上有改进。它只要一个场的存储，并按一定的方式自动启步，因此克服了菱形法作为三层格式共有的缺点，还能节约将近一半的工作量，此外，保留了显式和恒稳的优点，但也保留了相容性和精确度方面的缺点。

这种奇偶交替的算法思想很容易推广到高维以及其他类型的方程，程序实现也比较简单。

(4) 二维扩散方程

表 12.2 所列前五种都是一维格式的自然推广。我们知道，对于扩散方程，隐式是比较合适的。但在二维或三维时，隐式每步要解一个类似于椭圆型的差分方程，其系数阵不再是三对角线带状，没有象追赶法那样简便有效的解法，因此全隐式或平均或隐式虽有恒稳的优点，但使用上不方便。

交替方向法和局部一维法则是针对二维(或三维)特点而设计的。它们都是把每个时间步分解为两个(三维时为三个)分步。第一分步只在 x 方向采用隐式，这时在各横行上求解的代数方程组是彼此独立的，可以逐行用追赶法。类似地第二分步只在 y 方向用隐式。这样沿 x 和沿 y 互相交替，基本上是用解一维隐式的技巧来解高维问题。

在交替方向法中，对于方程

$$\frac{\partial u}{\partial t}-\left(b\frac{\partial^2 u}{\partial x^2}+b\frac{\partial^2 u}{\partial y^2}\right)-q=0 \tag{12.6.12}$$

在第一分步，$t_n\to t_{n+\frac{1}{2}}=t_n+\frac{\tau}{2}$，$\frac{\partial^2 u}{\partial x^2}$ 代以隐式差商，$\frac{\partial^2 u}{\partial y^2}$ 代以显式差商，得到

$$\frac{2}{\tau}(u_j^{n+\frac{1}{2}}-u_j^n)-\frac{b}{h_1^2}(\delta_x^2 u_{ij}^{n+\frac{1}{2}}+\delta_y^2 u_{ij}^n)-q_j^n=0 \tag{12.6.13}$$

在第二分步，$t_{n+\frac{1}{2}}\to t_{n+1}=t_{n+\frac{1}{2}}+\frac{\tau}{2}$，$\frac{\partial^2 u}{\partial y^2}$ 代以隐式差商而 $\frac{\partial^2 u}{\partial x^2}$ 代以显式差商，得到

$$\frac{2}{\tau}(u_j^{n+1}-u_j^{n+\frac{1}{2}})-\frac{b}{h_2^2}(\delta_x^2 u_{i,j}^{n+\frac{1}{2}}+\delta_y^2 u_{i,j}^{n+1})-q_j^{n+1}=0 \tag{12.6.14}$$

每步分步本身构成原方程(12.6.12)的一个相容逼近。如果单纯用第一分步的格式或第二分步的格式则各有增长因子

$$\lambda_1=\frac{1-\beta_1\sin^2\theta_1/2}{1+\beta_2\sin^2\theta_2/2},\quad \lambda_2=\frac{1-\beta_2\sin^2\theta_2/2}{1+\beta_1\sin^2\theta_1/2}$$

$$\beta_1=\tau b/h_1^2,\ \beta_2=\tau b/h_2^2,\ \theta_1=k_1h_1,\ \theta_2=k_2h_2$$

都只是条件稳。但交替使用时则增长因子 $\lambda=\lambda_1\lambda_2$ 相互补偿

$$\lambda=\frac{1-\beta_1\sin^2\theta_1/2}{1+\beta_2\sin^2\theta_2/2}\cdot\frac{1-\beta_2\sin^2\theta_2/2}{1+\beta_1\sin^2\theta_1/2}=\frac{1-\beta_1\sin^2\theta_1/2}{1+\beta_1\sin^2\theta_1/2}\cdot\frac{1-\beta_2\sin^2\theta_2/2}{1+\beta_2\sin^2\theta_2/2}$$

而 $|\lambda|\leqslant 1$，因此恒稳。也可以推广到三维，但公式比较复杂。

在局部一维法中，两个分步取为

$$\frac{1}{\tau}(u_j^{n+\frac{1}{2}}-u_j^n)-\frac{b}{h_1^2}\delta_x^2 u_{i,j}^{n+\frac{1}{2}}-\frac{1}{2}q_{i,j}^n=0 \qquad (12.6.15)$$

$$\frac{1}{\tau}(u_j^{n+1}-u_j^{n+\frac{1}{2}})-\frac{b}{h_2^2}\delta_y^2 u_{i,j}^{n+1}-\frac{1}{2}q_{i,j}^{n+1}=0 \qquad (12.6.16)$$

分别相当于微分方程：

$$\frac{1}{2}\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}-\frac{1}{2}q=0$$

$$\frac{1}{2}\frac{\partial u}{\partial t}-b\frac{\partial^2 u}{\partial x^2}-\frac{1}{2}q=0$$

因此每个分步并不构成原方程的相容逼近，仅当两步累加起来才与原方程相容。这里每一分步的公式就比较简单，并且各为恒稳，

$$\lambda_1=\frac{1}{1+4\beta_1\sin^2\theta_1/2},\quad \lambda_2=\frac{1}{1+4\beta_2\sin^2\theta_2/2}$$

因此合起来 $\lambda=\lambda_1\cdot\lambda_2$ 也是恒稳。这个方法很容易推广到高维以及其它类型的方程。

菱形法，特别是其改进形式即跳点法具有显式恒稳的优点，这项优点在二、三维扩散方程包括非线性在内更为显著，在程序实现上也比局部一维或交替方向法简单。因此，对于高维的，比较复杂而精度要求不太高的问题，菱形法和跳点法是可取的。

(5) 变系数扩散方程

表 12.1 及 12.2 是按照常系数扩散方程编排的。对于变系数方程，如

$$\frac{\partial}{\partial t}cu-\frac{\partial}{\partial x}b\frac{\partial u}{\partial x}-q=0$$

$$\frac{\partial}{\partial t}cu-\left(\frac{\partial}{\partial x}b\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}b\frac{\partial u}{\partial y}\right)-q=0$$

则引用时在公式中应作替换

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)\sim\frac{1}{\tau}(c_j^{n+1}u_j^{n+1}-c_j^n u_j^n)$$

$$\frac{1}{h^2}b(u_{j-1}^n-2u_j^n+u_{j+1}^n)\sim\frac{1}{h^2}[b_{j+\frac{1}{2}}^n(u_{j+1}^n-u_j^n)-b_{j-\frac{1}{2}}^n(u_j^n-u_{j-1}^n)]$$

$$\cdots$$

以保证守恒性(§12.5)。在介质系数 c, b, q 等有间断时，c_j^n, $b_{j+\frac{1}{2}}^n$, q_j^n 等应代以适当的平均值。在二维以及高维的情况下，当几何形状以及介质间断很复杂时，则以采用有限元法(第十四章)来离散化为宜。

§12.7 对流方程的差分格式

我们将对流方程的一些常用格式列在本节之末表 12.3，至于含时间 t 的三阶导数的波动方程的三层格式则列为表 12.4。将对流方程和扩散方程的格式适当地结合起来便得到对流-扩散方程的格式，列为表 12.5。对于一些有关的问题择要说明如下。

(1) 中心差格式及其改进

对于对流方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=q \qquad (12.7.1)$$

表 12.3 对流方程差分格式

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}-q=0$$

$$\alpha=\tau a/h,\ \theta=kh$$

名称	格式与截差	增长因子与稳定条件
右偏显式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_{j+1}^n-u_j^n)-q_j^n=0$ $E=O(\tau+h)$	$\lambda=1+\alpha-\alpha e^{i\theta}$ $-1\leqslant\alpha\leqslant0$
左偏显式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_j^n-u_{j-1}^n)-q_j^n=0$ $E=O(\tau+h)$	$\lambda=1-\alpha+\alpha e^{-i\theta}$ $0\leqslant\alpha\leqslant1$
右偏隐式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_{j+1}^{n+1}-u_j^{n+1})-q_j^{n+1}=0$ $E=O(\tau+h)$	$\lambda=(1-\alpha+\alpha e^{i\theta})^{-1}$ $\alpha\leqslant0$ 或 $1\leqslant\alpha$
左偏隐式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_j^{n+1}-u_{j-1}^{n+1})-q_j^{n+1}=0$ $E=O(\tau+h)$	$\lambda=(1+\alpha-\alpha e^{-i\theta})^{-1}$ $\alpha\leqslant-1$ 或 $0\leqslant\alpha$
中心差显式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-q_j^n=0$ $E=O(\tau+h^2)$	$\lambda=1-i\alpha\sin\theta$ 恒不稳
中心差隐式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})-q_j^{n+1}=0$ $E=O(\tau+h^2)$	$\lambda=(1+i\alpha\sin\theta)^{-1}$ 恒稳
平均隐式	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{1}{2}\left[\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-q_j^n\right]$ $+\frac{1}{2}\left[\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})-q_j^{n+1}\right]=0$ $E=O(\tau^2+h^2)$	$\lambda=\left(1-\frac{1}{2}i\alpha\sin\theta\right)\left(1+\frac{1}{2}i\alpha\sin\theta\right)^{-1}$ 恒稳
耗散中心差	$\frac{1}{\tau}\left[u_j^{n+1}-\frac{1}{2}(u_{j+1}^n+u_{j-1}^n)\right]+\frac{a}{2h}(u_{j-1}^n-2u_j^n+u_{j+1}^n)-q_j^n=0$ $E=O(h^2/\tau)+O(\tau+h)$	$\lambda=\cos\theta+i\alpha\sin\theta$ $\lvert\alpha\rvert\leqslant1$
三条腿 $q=0$	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-\frac{a^2\tau}{2h^2}(u_{j-1}^n-2u_j^n+u_{j+1}^n)=0$ $E=O(\tau^2+h^2)$	$\lambda=1-\alpha^2+\alpha^2\cos\theta-i\alpha\sin\theta$ $\lvert\alpha\rvert\leqslant1$
菱形	$\frac{1}{2\tau}(u_j^{n+2}-u_j^n)+\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})-q_j^{n+1}=0$ $E=O(\tau^2+h^2)$	$\lambda_{1,2}=i\alpha\sin\theta\pm\sqrt{1-\alpha^2\sin^2\theta}$ $\lvert\alpha\rvert\leqslant1$，此时 $\lvert\lambda\rvert=1$
跳点	$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)+q_j^n=0,\ n+1+j=$偶 $\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})+q_j^{n+1}=0,\ n+1+j=$奇 $E=O(\tau^2+h^2)$	同上

表 12.4 二阶波动方程差分格式

$$(\mathrm{I})\quad \frac{\partial^2 w}{\partial t^2}-a^2\frac{\partial^2 w}{\partial x^2}-q=0,\ a>0$$

$$(\mathrm{II})\quad \frac{\partial^2 w}{\partial t^2}-a^2\frac{\partial^2 w}{\partial x^2}+b\frac{\partial w}{\partial t}+cw=0,\ a>0;\ b,\ c>0$$

$$\delta_t^2 w_j^{n+1}=w_j^{n+2}-2w_j^{n+1}+w_j^n,\quad \delta_x^2 w_j^n=w_{j-1}^n-2w_j^n+w_{j-1}^n$$

$$\alpha=a\tau/h,\quad \beta=b\tau^2/h,\quad \gamma=c\tau^2,\quad \theta=kh$$

名称		格式与截差	增长因子与稳定条件
I	显式	$\frac{1}{\tau^2}\delta_t^2 w_j^{n+1}-\frac{a^2}{h^2}\delta_x^2 w_j^{n+1}-q_j^{n+1}=0$ $E=O(\tau^2+h^2)$	$\lambda^2+\left(4\alpha^2\sin^2\frac{\theta}{2}-2\right)\lambda+1=0$ $\alpha^2\leqslant 1$ 即 $\tau\leqslant h/a_n$
	全隐式	$\frac{1}{\tau^2}\delta_t^2 w_j^{n+1}-\frac{a^2}{h^2}\delta_x^2 w_j^{n+2}-q_j^{n+1}=0$ $E=O(\tau+h^2)$	$\left(1+4\alpha^2\sin^2\frac{\theta}{2}\right)\lambda^2-2\lambda+1=0$ 恒稳
	平均隐式	$\frac{1}{\tau^2}\delta_t^2 w_j^{n+1}-\frac{1}{2}\left[\frac{a^2}{h^2}\delta_x^2 w_j^n+\frac{a^2}{h^2}\delta_x^2 w_j^{n+2}\right]-q_j^{n+1}=0$ $E=O(\tau^2+h^2)$	$\left(1+2\alpha^2\sin^2\frac{\theta}{2}\right)\lambda^2-2\lambda+\left(1+2\alpha^2\sin^2\frac{\theta}{2}\right)=0$ 恒稳
II	显式	$\frac{1}{\tau^2}\delta_t^2 w_j^{n+1}-\frac{a^2}{h^2}\delta_x^2 w_j^{n+1}+\frac{b}{2h}(w_j^{n+2}-w_j^n)+cw_j^{n+1}-q_j^{n+1}=0$ $E=O(\tau^2+h^2)$	$\left(1+\frac{\beta}{2}\right)\lambda^2+\left(4\alpha^2\sin^2\frac{\theta}{2}-2+\gamma\right)\lambda+\left(1-\frac{\beta}{2}\right)=0$ $\alpha^2+\frac{\gamma}{4}\leqslant 1$ 即 $\tau\leqslant\frac{h}{\sqrt{a^2+ch^2/4}}$
	全隐式	$\frac{1}{\tau^2}\delta_t^2 w_j^{n+1}-\frac{a^2}{h^2}\delta_x^2 w_j^{n+2}+\frac{b}{2h}(w_j^{n+2}-w_j^n)+cw_j^{n+2}-q_j^{n+1}=0$ $E=O(\tau+h^2)$	$\left(1+4\alpha^2\sin^2\frac{\theta}{2}+\gamma+\frac{\beta}{2}\right)\lambda^2-2\lambda+\left(1-\frac{\beta}{2}\right)=0$ 恒稳
	平均隐式	$\frac{1}{\tau^2}\delta_t^2 w_j^{n+1}+\frac{1}{2}\left[\frac{a^2}{h^2}\delta_x^2 w_j^n+cw_j^n\right]+\frac{1}{2}\left[\frac{a^2}{h^2}\delta_x^2 w_j^{n+2}+cw_j^{n+2}\right]+\frac{b}{2h}[w_j^{n+2}-w_j^n]-q_j^{n+1}=0$ $E=O(\tau^2+h^2)$	$\left(1+2\alpha^2\sin^2\frac{\theta}{2}+\frac{\gamma}{2}+\frac{\beta}{2}\right)\lambda^2-2\lambda+\left(1+2\alpha^2\sin^2\frac{\theta}{2}+\frac{\gamma}{2}+\frac{\beta}{2}\right)=0$ 恒稳

表12.5 对流-扩散方程差分格式

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}-b\frac{\partial^2 u}{\partial x^2}-q=0,\ a\leqslant 0,\ b\geqslant 0$$

$$\delta^2u^n_j=u^n_{j-1}-2u^n_j+u^n_{j-1},\quad \alpha=a\tau/h,\quad \beta=b\tau/h^2,\quad \theta=kh$$

名称	格式与截差	增长因子与稳定条件
右偏显式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{h}(u^n_{j+1}-u^n_j)-\frac{b}{h^2}\delta^2u^n_j-q^n_j=0$ $E=O(\tau+h)$	$\lambda=1-(2\beta-\alpha)+(2\beta-\alpha)\cos\theta-i\alpha\sin\theta$ $0\leqslant 2\beta-\alpha\leqslant 1$ 并且 $\frac{\alpha^2}{\lvert 2\beta-\alpha\rvert}\leqslant 1$ 即 $2b-ah\geqslant 0$ 并且 $\tau\leqslant\min\left[\frac{h^2}{2h-ah},\ \frac{2b-ah}{a^2}\right]$
左偏显式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{h}(u^n_j-u^n_{j-1})+\frac{b}{h^2}\delta^2u^n_j-q^n_j=0$ $E=O(\tau+h)$	$\lambda=1-(2\beta+\alpha)+(2\beta+\alpha)\cos\theta-i\alpha\sin\theta$ $0\leqslant 2\beta+\alpha\leqslant 1$ 并且 $\frac{\alpha^2}{\lvert 2\beta+\alpha\rvert}\leqslant 1$ 即 $2b+ah\geqslant 0$ 并且 $\tau\leqslant\min\left[\frac{h^2}{2b+ah},\ \frac{2b+ah}{a^2}\right]$
右偏隐式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{h}(u^{n+1}_{j+1}-u^{n+1}_j)+\frac{b}{h^2}\delta^2u^{n+1}_j-q^{n+1}_j=0$ $E=O(\tau+h)$	$\lambda=[1+(2\beta-\alpha)-(2\beta-\alpha)\cos\theta+i\alpha\sin\theta]^{-1}$ $2\beta-\alpha\leqslant -1$ 或 $2\beta-\alpha\geqslant 0$
左偏隐式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{h}(u^{n+1}_j-u^{n+1}_{j-1})+\frac{b}{h^2}\delta^2u^{n+1}_j-q^{n+1}_j=0$ $E=O(\tau+h)$	$\lambda=[1+(2\beta+\alpha)-(2\beta+\alpha)\cos\theta+i\alpha\sin\theta]^{-1}$ $2\beta+\alpha\leqslant -1$ 或 $2\beta+\alpha\geqslant 0$
中心差显式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{2h}(u^n_{j+1}-u^n_{j-1})+\frac{b}{h^2}\delta^2u^n_j-q^n_j=0$ $E=O(\tau^2+h^2)$	$\lambda=1-2\beta+2\beta\cos\theta-i\alpha\sin\theta$ $2\beta\leqslant 1$ 并且 $\frac{\alpha^2}{2\beta}\leqslant 1$ 即 $\tau\leqslant\min\left[\frac{h^2}{2b},\ \frac{2b}{a^2}\right]$
全隐式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{2h}(u^{n+1}_{j+1}-u^{n+1}_{j-1})+\frac{b}{h^2}\delta^2u^{n+1}_j-q^n_j=0$ $E=O(\tau+h^2)$	$\lambda=(1+2\beta-2\beta\cos\theta+i\alpha\sin\theta)^{-1}$ 恒稳
平均隐式	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{1}{2}\left[\frac{a}{2h}(u^{n+1}_{j+1}-u^{n+1}_{j-1})+\frac{b}{h^2}\delta^2u^{n+1}_j-f^{n+1}_j\right]$ $+\frac{1}{2}\left[\frac{a}{2h}(u^n_{j+1}-u^n_{j-1})+\frac{b}{h^2}\delta^2u^n_j-f^n_j\right]=0$ $E=O(\tau^2+h^2)$	$\lambda=\left(1-\beta+\beta\cos\theta-\frac{i\alpha}{2}\sin\theta\right)$ $\cdot\left(1+\beta-\beta\cos\theta+\frac{i\alpha}{2}\sin\theta\right)^{-1}$ 恒稳
耗散中心差	$\frac{1}{\tau}\left[u^{n+1}_j-\frac{1}{2}(u^n_{j+1}+u^n_{j-1})\right]+\frac{a}{2h}(u^n_{j+1}-u^n_{j-1})-\frac{b}{h^2}\delta^2u^n_j-q^n_j=0$ $E=O(h^2/\tau)+O(\tau+h)$	$\lambda=-2\beta+(1+2\beta)\cos\theta-i\alpha\sin\theta$ 恒不稳
三条腿 $q=0$	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{2h}(u^n_{j+1}-u^n_{j+1})-\left(\frac{a^2\tau}{2h^2}+\frac{b}{h^2}\right)\delta^2u^n_j=0$ $E=O(\tau^2+h^2)$	$\lambda=1-(\alpha^2+2\beta)+(\alpha^2+2\beta)\cos\theta-i\alpha\sin\theta$ $\alpha^2+2\beta\leqslant 1$ 即 $a^2\tau^2+2b\tau\leqslant h^2$
菱形	$\frac{1}{2\tau}(u^{n+2}_j-u^n_j)+\frac{a}{2h}(u^{n+1}_{j+1}-u^{n+1}_{j-1})$ $-\frac{b}{h^2}(u^{n+1}_{j-1}-u^{n+2}_j-u^n_j+u^{n+1}_{j+1})-q^{n+1}_j=0$ $E=O(\tau^2/h^2)+O(\tau^2+h^2)$	$(1+2\beta)\lambda^2-(2\beta\cos\theta-i\alpha\sin\theta)\lambda$ $-(1-2\beta)=0$ $\lvert\alpha\rvert\leqslant 1$ 即 $\tau\leqslant h/\lvert a\rvert$
跳点	$\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{2h}(u^n_{j+1}-u^n_{j-1})-\frac{b}{h^2}\delta^2u^n_j-q^n_j=0,\ n+1+j=$偶 $\frac{1}{\tau}(u^{n+1}_j-u^n_j)+\frac{a}{2h}(u^{n+1}_{j+1}-u^{n+1}_{j-1})-\frac{b}{h^2}\delta^2u^{n+1}_j-q^{n+1}_j=0,\ n+1+j=$奇 $E=O(\tau^2/h^2)+O(\tau^2+h^2)$	同上

中心差格式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-q_j^n=0 \tag{12.7.2}$$

是一种最自然的格式，但是恒不稳，不能工作如把式中的 u_j^n 代以 $\frac{1}{2}(u_{j-1}^n+u_{j+1}^n)$ 则变为一种实用的格式

$$\frac{1}{\tau}\left[u_j^{n+1}-\frac{1}{2}(u_{j-1}^n+u_{j+1}^n)\right]+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-q_j^n=0 \tag{12.7.3}$$

叫做耗散中心差格式，或 Lax 格式。它的特征根

$$\lambda=\cos\theta+i\alpha\sin\theta,\quad \alpha=\frac{a\tau}{2h},\quad \theta=kh \tag{12.7.4}$$

的轨迹(当 θ 变化时)在复数平面内为以原点 O 为中心，横半径为 1，纵半径为 $|\alpha|$ 的椭圆，当这个椭圆含在单位圆之内时，格式稳定，其条件为

$$|\alpha|\leqslant 1,\quad 即\quad \tau\leqslant h/|a| \tag{12.7.5}$$

这就是库朗条件。

这个格式也可表成

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-\frac{1}{2\tau}(u_{j-1}^n-2u_j^n+u_{j+1}^n)-q_j^n=0 \tag{12.7.6}$$

可以解释为在不稳定的中心差公式(12.7.2)的基础上适当增加了一个起耗散作用的扩散项

$$-\frac{1}{2\tau}(u_{j-1}^n-2u_j^n+u_{j+1}^n)\approx-\left(\frac{h^2}{2\tau}\right)\frac{\partial^2 u}{\partial x^2} \tag{12.7.7}$$

因而提高了稳定性。因此称之为耗散中心差格式。

将(12.7.3)或(12.7.6)在节点 (x_j, t_n) 作幂次展开，得截断误差(相当于差分方程"减"微分方程)

$$E=-\left(\frac{h^2}{2\tau}\right)\left(\frac{\partial^2 u}{\partial x^2}\right)+O(\tau+h) \tag{12.7.8}$$

设 $h, \tau\to 0$ 并取 $\tau=O(h)$，则得到 $E=O(h)+O(\tau+h)=O(h)$，即具有一阶精度。但是，如果当 $h, \tau\to 0$ 时 $h^2/\tau=\sigma$ 保持为常值(例如取 $\tau=O(h^2)$)则由(12.7.8)知差分方程(12.7.3)的极限形式就不再是原来的对流方程(12.7.1)而是对流-扩散方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}-\frac{\sigma}{2}\frac{\partial^2 u}{\partial x^2}-q=0,\quad \sigma=h^2/\tau \tag{12.7.9}$$

多出了一个扩散项。试将(12.7.9)与(12.7.6～7)比较就更清楚。因此，在计算时，时间步长 τ 不能取得太小(相对于 h)，否则耗散太甚而失真。

对于 $q\equiv 0$ 即齐次方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0 \tag{12.7.10}$$

的情况，上述恒不稳的中心差公式(12.7.2)还可以改造成另一种很有效的格式。对(12.7.2)作幂次展开

$$\begin{aligned}0&=\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)\\&=\left(\frac{\partial u}{\partial t}\right)_j^n+\left(a\frac{\partial u}{\partial x}\right)_j^n+\frac{\tau}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_j^n+O(\tau^2+h^2)\end{aligned} \tag{12.7.11}$$

由于

$$\frac{\partial u}{\partial t}=-a\frac{\partial u}{\partial x},\ \frac{\partial^2 u}{\partial t}=-\left(\frac{\partial}{\partial t}a\frac{\partial u}{\partial x}\right)=-a\frac{\partial^2 u}{\partial t\partial x}=a^2\frac{\partial^2 u}{\partial x^2}$$

$$\left(\frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}\right)_j^n=\left(\frac{\tau a^2}{2}\frac{\partial^2 u}{\partial x^2}\right)_j^n=\frac{\tau a^2}{2h^2}(u_{j-1}^n-2u_j^n+u_{j+1}^n)+O(\tau h^2)$$

代入(12.7.11)右端即得所谓"三条腿"格式，也叫做 Lax-Wendroff 格式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j+1}^n)-\frac{a^2\tau}{2h^2}(u_{j-1}^n-2u_j^n+u_{j+1}^n)=0 \qquad (12.7.12)$$

具有二阶精度 $E=O(\tau^2+h^2)$。它的特征根

$$\lambda=1-\alpha^2+\alpha^2\cos\theta+i\alpha\sin\theta,\ \alpha=\tau a/h,\ \theta=kh \qquad (12.7.13)$$

其轨迹是以 $1-\alpha^2$ 为中心，横半径为 α^2，纵半径为 $|\alpha|$ 的椭圆，因此稳定条件为 $|\alpha|\leqslant 1$ 即 $\tau\leqslant h/|a|$，即库朗条件。与(12.7.2)比较，(12.7.12)多了一个扩散项

$$-\frac{a^2\tau}{2h^2}(u_{j-1}^n-2u_j^n+u_{j+1}^n)\approx-\left(\frac{a^2\tau}{2}\right)\frac{\partial^2 u}{\partial x^2} \qquad (12.7.14)$$

以提高稳定性。

这个格式可以化为分成两步走形式

$$\left.\begin{aligned}u_{j+\frac{1}{2}}^{n+\frac{1}{2}}&=\frac{1}{2}(u_{j+1}^n+u_j^n)-\frac{a\tau}{2h}(u_{j+1}^n-u_j^n)\\ u_j^{n+1}&=u_j^n-\frac{a\tau}{h}(u_{j+\frac{1}{2}}^{n+\frac{1}{2}}-u_{j-\frac{1}{2}}^{n+\frac{1}{2}})\end{aligned}\right\} \qquad (12.7.15)$$

前步是耗散中心差格式，后步是菱形格式(见下)。这种形式便于推广到拟线性的守恒型方程组。(12.7.12)和(12.7.15)是双曲型方程中应用较广的一种差分格式，它有较高的二阶精度，但只限于齐次方程(否则只有一阶精度)。

如果对方程(12.7.1)中的两个偏导数 $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial t}$ 都取中心差，则得"空心"的菱形格式

$$\left.\begin{aligned}\frac{1}{2\tau}(u_j^{n+2}-u_j^n)+\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})-q_j^n=0\\ E=O(\tau^2+h^2)\end{aligned}\right\} \qquad (12.7.16)$$

这是三层格式，特征方程是二次

$$\lambda^2+(i\alpha\sin\theta)\lambda-1=0,\ \lambda_{1,2}=i\alpha\sin\theta\pm\sqrt{1-\alpha^2\sin^2\theta} \qquad (12.7.17)$$

稳定条件是 $|\alpha|\leqslant 1$ 即库朗条件 $\tau\leqslant h/|a|$，这时有

$$|\lambda|\equiv 1 \qquad (12.7.18)$$

因此可以说差分格式是临界稳定或中立稳定。这与对流微分方程本身的临界稳定性(12.2.11)即 $\mathrm{Re}\mu=\mathrm{Re}iak=0$ 是契合的，因为差分方程的特征根 λ 与微分方程的特征根 μ 有对应关系 $\lambda\sim e^{\mu\tau}$。

菱形格式(12.7.16)虽有二阶精度，但为三层，使用不甚方便。但上面已经看到，它与耗散中心差格式相结合(引进半点)就变为两层格式而且保持了二阶精度。此外，与§12.6中(12.6.3)，(12.6.8~9)的情况相仿，菱形格式也等价于显式两层的跳点格式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)+q_j^n=0,\ n+1+j=\text{偶} \qquad (12.7.19)$$

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^{n+1}-u_{j-1}^{n+1})+q_j^{n+1}=0,\ n+1+j=\text{奇} \qquad (12.7.20)$$

即奇、偶、显、隐交替的方法，并且§12.6中的情况一样，还能改成计算上更节约的方式。这也是双曲型方程的常用方法之一。对于对流-扩散方程的推广见表12.5。值得指出，菱形或跳点法在扩散方程为恒稳，在对流方程为条件稳，服从库朗条件，而在扩散-对流方程仍为库朗条件稳，并不因增加扩散项而变苛刻，因此是比较有利的。

(2) 特征型差分格式

两个偏心格式(12.3.9～10)在实用上可以总结为

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\left\{\begin{array}{l}\frac{a}{h}(u_j^n-u_{j-1}^n)\\ \\ \frac{a}{h}(u_{j+1}^n-u_j^n)\end{array}\right\}-q_j^n=0,\quad 当\left\{\begin{array}{l}a\geqslant 0\\ a\leqslant 0\end{array}\right. \tag{12.7.21}$$

稳定条件

$$\tau\leqslant h/|a| \tag{12.7.22}$$

这就是说按 a 的正或负而取左偏或右偏，用以保证差分扰动与微分扰动沿相同方向传播，并且按照库朗条件取步长 τ。在几何上这表示格网三角形的斜边与特征线同倾向，而且从计算点 $(j,\ n+1)$ 向下引的特征线含在格网三角形内时稳定(图12.13)，在外时不稳定。

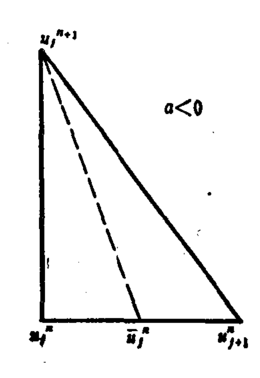


图　12.13

不难验证

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=\sqrt{1+a^2}\,\frac{\partial u}{\partial s}$$

这里 $\frac{\partial u}{\partial s}$ 表示向上沿特征线的方向导数。据此还可对特征型格式作另一解释，设 $a\geqslant 0$，由节点 $(j,\ n+1)$ 向下引特征线与第 n 层坐标线得一交点，位于节点 $(j,\ n)$ 的左方，它与节点 $(j,\ n+1)$ 的距离记为 Δs，u 在这点的值用 u_{j-1}^n 及 u_j^n 线性插出，记为 $\bar{u}_j^n$，

$$\Delta s=\sqrt{\tau^2+a^2\tau^2}=\tau\sqrt{1+a^2}$$

$$\bar{u}_j^n=\left(1-\frac{a\tau}{h}\right)u_j^n+\frac{a\tau}{h}u_{j-1}^n$$

注意当稳定即 $a\tau/h\leqslant 1$ 时为内插，否则为外插，将方向导数代为方向差商

$$\sqrt{1+a^2}\,\frac{\partial u}{\partial s}=\sqrt{1+a^2}\,\frac{1}{\Delta s}(u_j^{n+1}-u_{j-1}^n)=\frac{1}{\tau}(u_j^{n+1}-\bar{u}_j^n)$$

$$=\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_j^n-u_{j-1}^n)$$

因此特征型格式用线性插值方法产生沿特征线的方向导数的方法。

为了便于推广到方程组，公式(12.7.21)可以改写为统一的形式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{2h}(u_{j+1}^n-u_{j-1}^n)-\frac{|a|}{2h}(u_{j-1}^n-2u_j^n+u_{j+1}^n)-q_j^n=0 \qquad (12.7.23)$$

因此相当于在恒不稳的中心差显式(12.7.2)的基础上增加一个扩散项

$$-\frac{|a|}{2h}(u_{j-1}^n-2u_j^n+u_{j+1}^n)\approx-\left(\frac{|a|h}{2}\right)\frac{\partial^2 u}{\partial x^2} \qquad (12.7.24)$$

以提高稳定性。试与耗散中心差格式的(12.7.6)与三条腿格式的(12.7.12)相比较，大家都相当于对中心差格式增加一个起稳化作用的扩散项，只是扩散系数（是小参数）取得不同而已。

偏心格式也可以列为隐式：

左偏隐式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_j^{n+1}-u_{j-1}^{n+1})-q_j^{n+1}=0,\quad E=O(\tau+h) \qquad (12.7.25)$$

$$\lambda=(1+\alpha-\alpha e^{-i\theta})^{-1},\ \text{稳定条件}\quad \alpha\leqslant-1\quad \text{或}\quad 0\leqslant\alpha$$

右偏隐式

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\frac{a}{h}(u_{j+1}^{n+1}-u_j^{n+1})-q_j^{n+1}=0,\quad E=O(\tau+h) \qquad (12.7.26)$$

$$\lambda=(1-\alpha+\alpha e^{-i\theta})^{-1},\ \text{稳定条件}\quad \alpha\leqslant 0\quad \text{或}\quad 1\leqslant\alpha$$

对于(12.7.25)只取 $0\leqslant\alpha$，对于(12.7.26)只取 $\alpha\leqslant 0$，从而可以总结为特征型的恒稳隐式：

$$\frac{1}{\tau}(u_j^{n+1}-u_j^n)+\begin{Bmatrix}\frac{a}{h}(u_j^{n+1}-u_{j-1}^{n+1})\\ \\ \frac{a}{h}(u_{j+1}^{n+1}-u_j^{n+1})\end{Bmatrix}-f_j^{n+1}=0,\ \text{当}\begin{cases}a\geqslant 0\\ a\leqslant 0\end{cases} \qquad (12.7.27)$$

这个稳定的优点是实质上可以显式地定解。事实上，设 $a>0$，问题的边界条件应给在左端，比如说 $x=0$ 处，因此 u_0^{n+1} 已知，于是由公式(12.7.27)可逐次算出 u_1^{n+1}, u_2^{n+1}, …, 计算是自左至右。当 $a<0$ 时，则边界条件在右端而计算，自右至左。因此计算的方向恒与特征线的走向亦即扰动传播的方向一致。

(3) 边界处理

设定解域是 $0\leqslant x\leqslant X$, $0\leqslant t\leqslant T$

$$x_j=jh,\ j=0,1,\cdots,J;\ h=X/J$$
$$t_n=n\tau,\ n=0,1,\cdots,N;\ \tau=T/N$$

当 $a>0$ 时特征走向如图 12.13，在左边界 $x=0$ 要给定边界条件

$$u_0^{n+1}=\bar{u}_0^{n+1},\quad n=0,1,\cdots,N \qquad (12.7.28)$$

而右边界 $x=X$ 即 $j=J$ 处则不给条件，应由方程本身决定。无论采用怎样的差分格式，当从第 n 层推进到第 $n+1$ 层时在左边界 $j=0$ 处不按格式列差分方程而代以上列边界条件(12.7.28)。在右边界 $j=J$ 上，如果采用的中心差格式或其种种变形则要改用左偏即特征型的差分方程。当 $a<0$ 时情况类似，但反过来。特征型差分格式的一个优点就在于其对于边界处理的适应性，这是很自然的，因为边界条件本身总是与特征线相适应的(§12.2)。

(4) 显式和隐式的选择

根据 §12.4 以及表 12.3 可以知道，影响条件即库朗条件是显式波动差分方程的基本

判稳条件。对于波动过程而言，这是自然的条件。它对时间步长 τ 只要求与空间步长 h 同量级，$\tau=O(h)$，一般不算苛刻。因此,在多数情况没有必要采用隐式,这是与扩散过程有所不同的。但是，在有些问题中，例如弹性体的强迫振动，如果弹性体的尺寸比较小而使得 $\tau\approx h/|a|$ 比载荷波以持续时间 T 小若干量级，则库朗条件对于 τ 的限制也会造成沉重的负担，这时采用隐式就有好处。在对流与扩散并存的情况(见表12.5)由于有两种因素 $\alpha=\tau a/h$, $\beta=\tau b/h^2$ 的交互作用，判稳条件一般地要复杂些。一般说来对于显式，由于有了扩散的因素，对步长要求就比较严格，因此需要考虑隐式。值得指出，菱形格式及其改进的形式跳点格式在单纯对流方程时稳定条件是库朗条件，在单纯扩散方程时恒稳，而在对流——扩散方程中稳定条件仍保留为库朗条件，不因多了扩散项而变严。

§12.8 双曲型方程组

双曲型方程组，和其他类型的方程一样，通常导源于一组守恒律。取尤拉坐标下的一维可压缩气体的运动为例。刻划流场的有四个物理参数，密度 ρ，速度 u，单位质量的内能 e 和压力 p，其中 ρ, e, p 三者之间存在着一个函数关系，叫做状态方程，对于理想气体这就是

$$p=(\gamma-1)\rho e,\ \gamma\ 为绝热常数 \tag{12.8.1}$$

运动的基本规律是质量、动量和能量的守恒律。任取时段 $t'\leqslant t\leqslant t''$ 及空间区段 $x'\leqslant x\leqslant x''$，质量守恒律表为

$$\int_{x'}^{x''}(\rho)_{t=t''}dx-\int_{x'}^{x''}(\rho)_{t=t'}dx=\int_{t'}^{t''}(\rho u)_{x=x'}dt-\int_{t'}^{t''}(\rho u)_{x=x''}dt \tag{12.8.2}$$

表示$[x', x'']$内质量的增加是通过边界 $x=x'$ 及 $x=x''$ 顺流带进的。动量守恒律为

$$\int_{x'}^{x''}(\rho u)_{t=t''}dx-\int_{x'}^{x''}(\rho u)_{t=t'}dx=\int_{t'}^{t''}(\rho u\cdot u+p)_{x=x'}dt-\int_{t'}^{t''}(\rho u\cdot u+p)_{x=x''}dt \tag{12.8.3}$$

表示动量的增加是由于顺流带进了动量以及两端压差引起的。能量(内能+动能)守恒律为

$$\begin{aligned}&\int_{x'}^{x''}\rho\left(e+\frac{1}{2}u^2\right)_{t=t''}dx-\int_{x'}^{x''}\rho\left(e+\frac{1}{2}u^2\right)_{t=t'}dx\\&=\int_{t'}^{t''}\left[\rho\left(e+\frac{1}{2}u^2\right)u+pu\right]_{x=x'}dt-\int_{t'}^{t''}\left[\rho\left(e+\frac{1}{2}u^2\right)u+pu\right]_{x=x''}dt\end{aligned} \tag{12.8.4}$$

表示能量增加是由于顺流带进了能量以及两端的压力功差引起的。假定场量有一定的光滑性。于是命 t', $t''\to t$, 上述三式可以化为

$$\frac{\partial}{\partial t}\int_{x'}^{x''}\rho dx=(\rho u)_{x=x'}-(\rho u)_{x=x''}$$

$$\frac{\partial}{\partial t}\int_{x'}^{x''}\rho u dx=(\rho u^2+p)_{x=x'}-(\rho u^2+p)_{x=x''}$$

$$\frac{\partial}{\partial t}\int_{x'}^{x''}\rho\left(e+\frac{1}{2}u^2\right)dx=\left[\rho\left(e+\frac{1}{2}u^2\right)pu\right]_{x=x'}-\left[\rho\left(e+\frac{1}{2}u^2\right)u+pu\right]_{x=x''}$$

再命 x, $x''\to x$ 则得

$$\frac{\partial}{\partial t}\rho+\frac{\partial}{\partial x}\rho u=0 \tag{12.8.5}$$

$$\frac{\partial}{\partial t}\rho u+\frac{\partial}{\partial x}(\rho u^2+p)=0 \tag{12.8.6}$$

$$\frac{\partial}{\partial t}\rho\left(e+\frac{1}{2}u^2\right)+\frac{\partial}{\partial x}\left[\rho\left(e+\frac{1}{2}u^2\right)u+pu\right]=0 \tag{12.8.7}$$

这就是微分形式的守恒律,连同(12.8.1)共四个方程以定四个量 ρ, u, e, p。

在一般情况,设有

$$\boldsymbol{U}=\begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \quad \boldsymbol{\Phi}(\boldsymbol{U})=\begin{bmatrix} \varphi_1(U) \\ \vdots \\ \varphi_m(U) \end{bmatrix}, \quad \boldsymbol{Q}=\begin{bmatrix} q_1 \\ \vdots \\ q_m \end{bmatrix}$$

这里 u_1, …, u_m 是一组守恒的物理量, $\varphi_1(u_1, \cdots, u_m)$, …, $\varphi_m(u_1, \cdots, u_m)$ 是 u_1, …, u_m 的一组已知函数, q_1, …, q_m 为一组分别对应 u_1, …, u_m 的源项。守恒律的积分形式是

$$\int_{x'}^{x''}(\boldsymbol{U})_{t=t''}dx-\int_{x'}^{x''}(\boldsymbol{U})_{t=t'}dx$$
$$=\int_{t'}^{t''}(\boldsymbol{\Phi}(\boldsymbol{U}))_{x=x'}dt-\int_{t'}^{t''}(\boldsymbol{\Phi}(\boldsymbol{U}))_{x=x''}dt+\int_{t'}^{t''}\int_{x'}^{x''}\boldsymbol{Q}\,dx\,dt \tag{12.8.8}$$

或者

$$\frac{\partial}{\partial t}\int_{x'}^{x''}\boldsymbol{U}dx=\boldsymbol{\Phi}(\boldsymbol{U})_{x=x'}-\boldsymbol{\Phi}(\boldsymbol{U})_{x=x''}+\int_{x'}^{x''}\boldsymbol{Q}dx \tag{12.8.9}$$

从而得守恒型的微分方程

$$\frac{\partial \boldsymbol{U}}{\partial t}+\frac{\partial \boldsymbol{\Phi}(\boldsymbol{U})}{\partial x}=\boldsymbol{Q} \tag{12.8.10}$$

在前面的例子中,

$$\boldsymbol{U}=\begin{bmatrix} \rho \\ \rho u \\ \rho\left(e+\frac{1}{2}u^2\right) \end{bmatrix}, \quad \boldsymbol{\Phi}=\begin{bmatrix} \rho u \\ \rho u^2+p \\ \rho\left(e+\frac{1}{2}u^2\right)u+pu \end{bmatrix}, \quad \boldsymbol{Q}=0, \quad \boldsymbol{p}=(\gamma-1)\rho e \tag{12.8.11}$$

根据微分公式

$$\frac{\partial}{\partial x}\varphi_i(u_1, \cdots, u_m)=\sum_{j=1}^{m}\frac{\partial \varphi_i}{\partial u_j}\frac{\partial u_j}{\partial x}$$

方程组(12.8.10)可以化为通常所谓的标准型

$$\frac{\partial \boldsymbol{U}}{\partial t}+\boldsymbol{A}\frac{\partial \boldsymbol{U}}{\partial x}=\boldsymbol{Q} \tag{12.8.12}$$

这里 $\boldsymbol{A}$ 就是函数组 φ_1, …, φ_m 对于 u_1, …, u_m 的导数矩阵即雅谷比矩阵

$$\boldsymbol{A}=[a_{ij}]=\left[\frac{\partial \varphi_i}{\partial u_j}\right] \tag{12.8.13}$$

方程组(12.8.11)叫做拟线性的,如果矩阵 $\boldsymbol{A}$ 的元依赖于 u_1, …, u_m, 即 $\boldsymbol{A}=\boldsymbol{A}(\boldsymbol{U})$, 反之,如果 $\boldsymbol{A}$ 的元不依赖 $\boldsymbol{U}$, 则方程组叫做线性的。

当矩阵 $\boldsymbol{A}$ 的元依赖于 u_1, …, u_m 即 $\boldsymbol{A}=\boldsymbol{A}(\boldsymbol{U})$ 时, 称方程组(12.8.12)为拟线性的, 拟线性方程是非线性方程中比较简单而实践上最重要的一种。(12.8.5~7)就是拟线性的。当 $\boldsymbol{A}$ 的元不依赖于 u_1, …, u_m 时则称方程组为线性的, 这时如果 $\boldsymbol{A}$ 的元依赖于 x, t 即 $\boldsymbol{A}=\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{t})$, 则称方程组为变系数的,否则则称为常系数的。方程(12.1.1)和(12.1.8)是常系数的。

当矩阵 $\boldsymbol{A}$ 的本征值 a_1, …, a_m 都是实数而且有线性无关的本征向量组, 则称方程组(12.8.12)为双曲型的。这时取矩阵 $\boldsymbol{P}=[p_{ij}]$, 它的行向量是 $\boldsymbol{A}$ 的本征行向量,于是

$$\boldsymbol{PA}=\boldsymbol{\Lambda P},\ \boldsymbol{\Lambda}=\begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_m \end{bmatrix} \tag{12.8.14}$$

当 $\boldsymbol{A}$ 依赖于 $\boldsymbol{U}$ 时，它的本征值 a_i 和本征行向量也依赖于 $\boldsymbol{U}$。用矩阵 $\boldsymbol{P}$ 左乘(12.8.12)的两端，并利用(12.8.14)可得

$$\boldsymbol{P}\frac{\partial \boldsymbol{U}}{\partial t}+\boldsymbol{\Lambda P}\frac{\partial \boldsymbol{U}}{\partial x}=\boldsymbol{PQ} \tag{12.8.15}$$

或写成分量的形式

$$\sum_{j=1}^{m} p_{ij}\left(\frac{\partial u_j}{\partial t}+a_i\frac{\partial u_j}{\partial x}-q_j\right)=0,\ i=1,\ \cdots,\ m \tag{12.8.16}$$

这就是方程组(12.8.10)的特征型，只有双曲型方程组才能化成特征型。注意(12.8.16)与单向波方程非常相似。在单向波方程通过每个点 $(x,\ t)$ 有一条特征线 $x-at=\text{const}$，即斜率为 a 的直线。在(12.8.16)通过每个点 $(x,\ t)$ 有 m 根曲线，也叫做特征线，其斜率分为

$$\frac{dx}{dt}=a_i,\ i=1,\ \cdots,\ m$$

总的说来共有 m 族特征线。在拟线性或变系数时一般都是曲线，在常系数时特征线是直线。(12.8.16)中每个方程中的导数只以一个特征线方向的方向导数 $\left(\frac{\partial}{\partial t}+a_i\frac{\partial}{\partial x}\right)$ 的形式出现。因此，如命第 i 特征线的弧长坐标为 s_i，相应的方向微商为 $\frac{\partial}{\partial s_i}$ 则(12.8.16)也可以表为

$$\sum_{j=1}^{m} p_{ij}\left(\sqrt{1+a_i}\frac{\partial u_j}{\partial s_i}-q_j\right)=0,\ i=1,\ \cdots,\ m \tag{12.8.17}$$

每个方程表示沿一条特征线的增量关系式，共有 m 个“特征关系式”。

可以证明，方程组(12.8.10)的解的小扰动在 x，t 平面上是沿着 m 族特征线传播的，也可以说小扰动以特征速度 a_1，$\cdots$，a_m 传播，当 $a_i>0$ 为向 $+x$ 方向传播，$a_i<0$ 时则向 $-x$ 方向传播。

根据特征线的概念可以明确边界条件的给法(§12.2)。设定解区域为

$$0\leqslant x\leqslant X,\ 0\leqslant t\leqslant T$$

除了初始条件即给定 $u_1(x,\ 0)$，$\cdots$，$u_m(x,\ 0)$ 外，还要规定边界条件。在左边界 $(0,\ t)$ 上，设 a_i 中有 p 个为正，即向下引的特征线有 p 条指向界外，则相应的 p 个特征关系失效，这时应给定 p 个边界条件以补足之。在右边界 $(X,\ t)$ 上，设 a_i 中有 q 个为负，即向下引的特征线有 q 条指向界外，则相应 q 个特征关系失效而应代以 q 个边界条件，$p+q\leqslant m$。

图 12.14

对于一维气体运动方程(12.8.5～7)，不难写出其标准型和特征型。实践上往往改取 ρ，u，p 为独立因变量，经简化得标准型

$$\frac{\partial \boldsymbol{U}}{\partial t}+\boldsymbol{A}\frac{\partial \boldsymbol{U}}{\partial x}=0$$

$$\boldsymbol{U}=\begin{bmatrix}\rho\\u\\p\end{bmatrix},\qquad \boldsymbol{A}=\begin{bmatrix}u&\rho&0\\0&u&\dfrac{1}{\rho}\\0&\rho c^2&u\end{bmatrix},\qquad c=\sqrt{\gamma p/\rho}=\text{声速}\tag{12.8.18}$$

$\boldsymbol{A}$ 的本征值为 $u-c,\ u,\ u+c$，即

$$\boldsymbol{\Lambda}=\begin{bmatrix}u-c&0&0\\0&u&0\\0&0&u+c\end{bmatrix},\qquad \boldsymbol{P}=\begin{bmatrix}0&-\rho c&1\\-c^2&0&1\\0&\rho c&1\end{bmatrix}\tag{12.8.19}$$

因此特征型表为

$$\left.\begin{aligned}&-\rho c\Big(\frac{\partial}{\partial t}+(u-c)\frac{\partial}{\partial x}\Big)u+\Big(\frac{\partial}{\partial t}+(u-c)\frac{\partial}{\partial x}\Big)p=0\\&-c^2\Big(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}\Big)\rho+\Big(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}\Big)p=0\\&\rho c\Big(\frac{\partial}{\partial t}+(u+c)\frac{\partial}{\partial x}\Big)u+\Big(\frac{\partial}{\partial t}+(u+c)\frac{\partial}{\partial x}\Big)p=0\end{aligned}\right\}\tag{12.8.20}$$

§ 12.9 双曲型方程组的差分格式

对流即单向波方程的差分格式都可以适当地推广到双曲型方程组或守恒的方程组。

(1) 特征型差分格式

双曲方程组有一类常用的数值解法是所谓特征线法，它是在特征线和特征关系的基础上进行的，在解的推进过程中逐步构造特征线网，一般说来形状是不规则的。这类方法历史发展比较早，可以达到较高的精度，但逻辑上比较复杂，对此将不作介绍，可以参考专门的著作如[2]。下面的方法可以看做特征线法的一种变形和简化。

在将方程组表为特征型

$$\sum_{k=1}^{m}p_{ik}\left[\frac{\partial u_k}{\partial t}+a_i\frac{\partial u_k}{\partial x}-q_{ik}\right]=0,\quad i=1,\ \cdots,\ m\tag{12.9.1}$$

的基础上，对其中每个方程运用特征型偏心格式，即

$$\sum_{k=1}^{m}p_{ik}\left[\frac{1}{\tau}(u_{k,j}^{n+1}-u_{k,j}^{n})+\frac{a_k}{h}\delta_k u_{k,j}^{n}-q_{k,j}\right]=0,\ i=1,\ \cdots,\ m\tag{12.9.2}$$

这里 $u_{k,j}^{n}$ 表示 u_k 在节点 $(x_j,\ t_n)$ 处的值，而空间差分按 a_i 的正负取左偏或右偏

$$\frac{a_k}{h}\delta_k u_{k,j}^{n}=\begin{cases}\dfrac{a_k}{h}(u_{k,j}^{n}-u_{k,j-1}^{n}),\quad \text{当}\quad a_k\geqslant 0\\[2ex]\dfrac{a_k}{h}(u_{k,j+1}^{n}-u_{k,j}^{n}),\quad \text{当}\quad a_k<0\end{cases}$$

于是在每个节点 $(x_j,\ t_{n+1})$ 得到含有未知数 $u_{1,j}^{n+1},\ \cdots,\ u_{m,j}^{n+1}$ 的方程组需要解算：

$$\sum_{k=1}^{m}p_{ik}u_{k,j}^{n+1}=\sum_{k=1}^{m}p_{ik}\Big(u_{k,j}^{n}-\frac{\tau a_i}{h}\delta_k u_{k,j}^{n}+\tau q_{k,j}\Big)\tag{12.9.3}$$

当 p_{ik} 依赖于 U 即非线性时可以取 $p_{ik}=p_{ik}(U^n)$ 库朗条件则推广为：

$$\tau\leqslant\frac{h}{\max|a_i|}\quad\text{或}\quad\tau<\frac{h}{\max|a_i|}\tag{12.9.4}$$

实践上多采用严格不等式的形式。在变系数或非线性的情况，这个条件必须逐点满足，故随着时间的推进 τ 可能经常变化。

在边界点上，根据 a_i 的正负号，必须剔除失效的特征方程，而将余下的方程与给定的边界条件(加起来还是 m 个方程)联解。

这是一个有效的差分格式，对于非线性方程组乃至有间断解的情况也是适用的。不便之处在于需要预先把方程组化为特征型，每点还要联解一个方程组。为了改进，可以采用等价的形式(12.7.7)，即先将(12.9.2)写成

$$\sum_{k=1}^{m} p_{ik,j}^{n}\Big[\frac{1}{\tau}(u_{k,j}^{n+1}-u_{k,j}^{n})+\frac{a_k}{2h}(u_{k,j+1}^{n}-u_{k,j-1}^{n})$$
$$-\frac{|a_k|}{2h}(u_{k,j-1}^{n}-2u_{k,j}^{n}+u_{k,j+1}^{n})-q_{k,j}\Big]=0 \tag{12.9.5}$$
$$E=O(\tau+h)$$

命

$$\boldsymbol{\Lambda}=\begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{bmatrix}, \qquad \widetilde{\boldsymbol{\Lambda}}=\begin{bmatrix} |a_1| & & 0 \\ & \ddots & \\ 0 & & |a_n| \end{bmatrix}$$

于是

$$P_j^n\frac{1}{\tau}(U_j^{n+1}-U_j^n)+\Lambda_j^nP_j^n\frac{1}{2h}(U_{j+1}^n-U_{j-1}^n)-\widetilde{\Lambda}_j^nP_j^n\frac{1}{2h}(U_{j-1}^n-2U_j^n+U_{j+1}^n)-P_j^nQ_j^n=0$$

用 $\boldsymbol{P}^{-1}$ 左乘两端，注意

$$\boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}=\boldsymbol{A}$$

并且命

$$\widetilde{\boldsymbol{A}}=\boldsymbol{P}^{-1}\widetilde{\boldsymbol{\Lambda}}\boldsymbol{P} \tag{12.9.6}$$

便得到等价的特征型格式

$$\frac{1}{\tau}(U_j^{n+1}-U_j^n)+\frac{1}{2h}A_j^n(U_{j+1}^n-U_{j-1}^n)$$
$$-\frac{1}{2h}\widetilde{A}_j^n(U_{j-1}^n-2U_j^n+U_{j+1}^n)-Q_j^n=0 \tag{12.9.7}$$

这里矩阵 $\widetilde{\boldsymbol{A}}$ 即(12.9.6)可以表为阵 $\boldsymbol{A}$ 的幂次和

$$\widetilde{\boldsymbol{A}}=\sum_{p=0}^{m-1}\alpha_p\boldsymbol{A}^p$$

系数 $\alpha_0, \cdots, \alpha_{m-1}$ 是线代数方程组

$$\sum_{p=0}^{m-1}\alpha_p a_i^p=|a_i|,\quad i=1, \cdots, m$$

的解，详见[3]。

(2) 守恒型差分格式

可以仿照§12.5的方法在(12.8.7)的基础上构造守恒型的差分格式。取格网同于§12.5，命

$$C_j^n=\int_{x_{j-\frac12}}^{x_{j+\frac12}}U_{t=t_n}dx,\quad j=0, 1, \cdots, J;\ n=0, 1, \cdots, N$$

$$B_{j+\frac12}^{n+\frac12}=\int_{t_n}^{t_{n+1}}\Phi_{x=x_{j+\frac12}}dt,\quad j=0, 1, \cdots, J;\ n=0, 1, \cdots, N-1$$

$$Q_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\int_{t_n}^{t_{n+1}}dt\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}Qdx,\ \ j=0,\ 1,\ \cdots,\ J;\ \ n=0,\ 1,\ \cdots,\ N-1$$

于是有

$$C_j^{n+1}-C_j^n+B_{j+\frac{1}{2}}^{n+\frac{1}{2}}-B_{j-\frac{1}{2}}^{n+\frac{1}{2}}+Q_j^{n+\frac{1}{2}}=0,\ \ j=0,\ 1,\ \cdots,\ J;\ n=0,\ 1,\ \cdots,\ N \tag{12.9.8}$$

可以取，比方说

$$C_j^n\approx U_j^n h_j$$

$$B_{j+\frac{1}{2}}^{n+\frac{1}{2}}\approx\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}\tau_{n+\frac{1}{2}}$$

$$F_j^{n+\frac{1}{2}}\approx\overline{Q}_j^{n+\frac{1}{2}}h_j\tau_{n+\frac{1}{2}}$$

这里 $\overline{\Phi}$, $\overline{Q}$ 表示相应线段或面元上的某种均值，具体尚待定。于是形式上有

$$(U_j^{n+1}-U_j^n)h_j+(\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}-\overline{\Phi}_{j-\frac{1}{2}}^{n+\frac{1}{2}})\tau_{n+\frac{1}{2}}+\overline{Q}_j^{n+\frac{1}{2}}h_j\tau_{n+\frac{1}{2}}=0 \tag{12.9.9}$$

按照 $\overline{\Phi}$, $\overline{Q}$ 的取法，可以把对流方程的一些主要格式推广为守恒型(取等距格网)

(i) 中心差显式：取

$$\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\frac{1}{2}[\Phi(U_j^n)+\Phi(U_{j+1}^n)]$$

$$\overline{Q}_j^{n+\frac{1}{2}}=Q_j^n$$

得到

$$h(U_j^{n+1}-U_j^n)+\frac{\tau}{2}[\Phi(U_{j+1}^n)-\Phi(U_{j-1}^n)]-Q_j^n=0 \tag{12.9.10}$$

$$E=O(\tau+h^2),\ \text{恒不稳}$$

(ii) 中心差隐式：取

$$\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\frac{1}{2}[\Phi(U_j^{n+1})+\Phi(U_{j+1}^{n+1})]$$

$$\overline{Q}_j^{n+\frac{1}{2}}=Q_j^{n+1}$$

得到

$$h(U_j^{n+1}-U_j^n)+\frac{\tau}{2}[\Phi(U_{j+1}^{n+1})-\Phi(U_{j-1}^{n+1})]-Q_j^{n+1}=0 \tag{12.9.11}$$

$$E=O(\tau+h^2),\ \text{恒稳}$$

(iii) 耗散中心差：取

$$\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\frac{1}{2}[\Phi(U_{j+1}^n)+\Phi(U_{j-1}^n)]-\frac{h}{2\tau}(U_{j+1}^n-U_{j-1}^n)$$

$$\overline{Q}_{j+\frac{1}{2}}^{n+\frac{1}{2}}=Q_j^n$$

得到

$$h[U_j^{n+1}-U_j^n]+\frac{\tau}{2}[\Phi(U_{j+1}^n)-\Phi(U_{j-1}^n)]-Q_j^n=0 \tag{12.9.12}$$

相当于(12.7.2)，稳定条件为(12.9.4)。

(iv) 三条腿格式——对于 $Q=0$ 的情况：取

$$U_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\frac{1}{2}(U_j^n+U_{j+1}^n)+\frac{\tau}{2h}[\Phi(U_{j+1}^n)-\Phi(U_j^n)]$$

$$\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\Phi(U_{j+\frac{1}{2}}^{n+\frac{1}{2}})$$

得到

$$h(U_j^{n+1}-U_j^n)+\tau[\Phi(U_{j+\frac{1}{2}}^{n+\frac{1}{2}})-\Phi(U_{j-\frac{1}{2}}^{n+\frac{1}{2}})]=0 \tag{12.9.13}$$

$$E=O(\tau^2+h^2)$$

这就是分两步走的格式(12.7.7),稳定条件同(12.9.4)。

(v) 守恒的特征型差分格式: 取

$$\overline{\Phi}_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\frac{1}{2}[\Phi(U_j^n)+\Phi(U_{j+1}^n)]-\hat{A}_{j+\frac{1}{2}}^n(U_{j+1}^n-U_j^n)$$

$$\overline{Q}_j^{n+\frac{1}{2}}=Q_j^n$$

$$\hat{A}_{j+\frac{1}{2}}^n=\frac{1}{2}[A(U_j^n)+A(U_{j+1}^n)]$$

得到

$$\begin{aligned}&h(U_j^{n+1}-U_j^n)+\frac{\tau}{2}[\Phi(U_{j+1}^n)-\Phi(U_{j-1}^n)]\\&-\tau[A_{j+\frac{1}{2}}^n(U_{j+1}^n-U_j^n)-A_{j-\frac{1}{2}}^n(U_j^n-U_{j-1}^n)]-Q_j^n=0\end{aligned} \tag{12.9.14}$$

相当于(12.9.7),稳定条件同(12.9.4)。

(3) 交错格网

对于含有多个未知函数的方程,还可以构造交错格网的差分格式。例如,对于二阶的波动方程

$$\frac{\partial^2 w}{\partial t^2}-a^2\frac{\partial^2 w}{\partial x^2}=f,\quad a>0$$

命 $u=\frac{\partial w}{\partial t}$, $v=a\frac{\partial w}{\partial x}$,得等价的一阶方程组

$$\left.\begin{aligned}&\frac{\partial u}{\partial t}-a\frac{\partial v}{\partial x}=f\\&\frac{\partial v}{\partial t}-a\frac{\partial u}{\partial x}=g=0\end{aligned}\right\} \tag{12.9.15}$$

图 12.15

根据这个特殊形式,可以把 u, v, f, g 配置于不同类型的节点上如 $u_{j+\frac{1}{2}}^n$, $v_j^{n+\frac{1}{2}}$, f_j^n, $g_{j+\frac{1}{2}}^{n+\frac{1}{2}}$, 如图 12.15。然后对两个方程分别运用菱形中心差格式而得到

$$\left.\begin{aligned}&\frac{1}{\tau}(u_{j+\frac{1}{2}}^{n+1}-u_{j+\frac{1}{2}}^n)-\frac{a}{h}(v_{j+1}^{n+\frac{1}{2}}-v_j^{n+\frac{1}{2}})-g_{j+\frac{1}{2}}^{n+\frac{1}{2}}=0\\&\frac{1}{\tau}(v_j^{n+\frac{1}{2}+1}-v_j^{n+\frac{1}{2}})-\frac{a}{h}(u_{j+\frac{1}{2}}^{n+1}-u_{j-\frac{1}{2}}^{n+1})-f_j^{n+1}=0\end{aligned}\right\} \tag{12.9.16}$$

$$E=O(\tau^2+h^2)$$

特征方程是

$$\lambda^2+(4\alpha^2\sin^2\theta/2-2)\lambda+1=0$$

稳定条件是

$$\alpha^2\leqslant 1 \quad 即 \quad \tau\leqslant h/a$$

当然也可以把 u, v 等均置于整点 (x_j, t_n) 而直接运用菱形公式,具有同阶的精度与相同的稳定性能,但中心差是跨点进行的间距为 2τ, $2h$。(12.9.16) 的节点是交错配置适宜中心差的间距成为 τ, h,缩小了一半,实际上提高了精度。

(4) 边界处理

对于方程组，边界处理往往是很麻烦的问题。一种处理方法是不论内点用什么格式，边界点恒采用特征格式与边界条件相结合的形式。在左边界上，设特征值 $a_1, \cdots, a_m$ 中有 p 个 >0，则应有 p 个边界条件 $\varphi_1(u_1, \cdots, u_m)=0, \cdots, \varphi_p(u_1, \cdots, u_m)=0$，从特征型方程组中剔去相应的 p 个，对余下的 $m-p$ 个特征型方程（对应于 $a_i \geqslant 0$）按照特征型即右偏格式列出 $m-p$ 个差分方程，加上 p 个边界条件联解出 m 个未知数。在右边界上，设 $a_1, \cdots, a_m$ 中有 q 个 <0 则应有 q 个边界条件。这时从特征型方程组中剔去相应的 q 个方程，对余下的 $m-q$ 个方程采用特征型——这时为左偏——格式再与边界条件联解。

(5) 关于间断解的计算

拟线性双曲型方程可能有间断解，即使初始条件是光滑的，也有可能在随后的时间内自发地形成间断解，如气体力学中的冲击波。在间断点，解不再适合原始的微分方程而是满足所谓间断条件。以方程组(12.8.5～7)为例，间断条件则为

$$\rho^+ \frac{dx}{dt} + (\rho u)^+ = \rho^- \frac{dx}{dt} + (\rho u)^- \tag{12.9.17}$$

$$(\rho u)^+ \frac{dx}{dt} + (\rho u^2+p)^+ = (\rho u)^- \frac{dx}{dt} + (\rho u^2+p)^- \tag{12.9.18}$$

$$\begin{aligned} &\left[\rho\left(e+\frac{1}{2}u^2\right)\right]^+ \frac{dx}{dt} + \left[\rho\left(e+\frac{1}{2}u^2\right)u+pu\right]^+ \\ &= \left[\rho\left(e+\frac{1}{2}u^2\right)\right]^- \frac{dx}{dt} + \left[\rho\left(e+\frac{1}{2}u^2\right)u+pu\right]^- \end{aligned} \tag{12.9.19}$$

$\dfrac{dx}{dt}$ 表示间断点移动的速度即击波速度。方程 (12.9.17～19) 实质上是守恒律在间断点的表达形式，作差分形式。它光滑点上的微分形式的守恒律(12.8.5～7)一样都是从积分守恒律(12.8.2～4)导出的。

对于间断性有两类处理方法。一类是间断分离法，即对间断点进行跟踪，把它隔离出来单独处理，要求满足间断条件如(12.9.17～19)。这类方法精度较高，但逻辑上比较复杂，特别当物理条件复杂以及二维或高维的情况则相当困难。关于在特征线法的基础上间断跟踪的处理见[2]。另一类所谓穿行计算法，即不论有无间断，不加区别恒按统一的格式处理。这时间断解被表为具有一定过渡层的连续解（如图 12.16），但后者应能基本上反映本来的间断特征。事实上不是随便的差分格式都能做到这一点的。为此目的，通常在待解的微分方程组中外加一定的扩散项如 $\varepsilon \dfrac{\partial^2 u}{\partial x^2}$ 或更确切一些如 $\dfrac{\partial}{\partial x} \varepsilon \dfrac{\partial u}{\partial x}$ 在冲击波计算中，这种人为引进的项起着粘性的作用，因此叫做人工粘性项，产生平滑化的效果，把间断解平滑化为具有过渡层的连续解。ε 是一个小参数，应随 $\tau, h \to 0$ 而 $\to 0$。顺便指出，本章介绍的一些

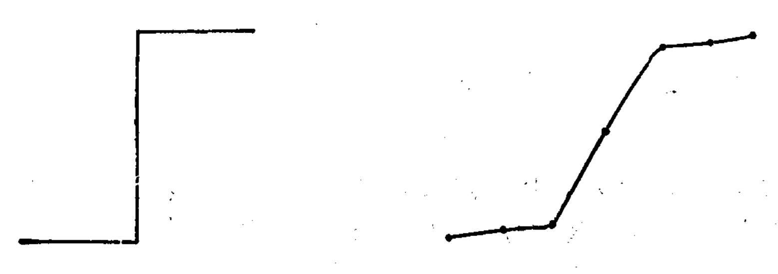


图 12.16

差分格式中，虽然没有事先明显引入粘性项，但差分格式的构造中自动蕴藏了这样的项，如耗散中心差格式(12.7.2)相当于取 $\varepsilon=\frac{h^2}{2\tau}$三条腿格式(12.7.3)相当于取 $\varepsilon=\frac{\tau a^2}{2}$，特征型格式(12.7.7)相当于取 $\varepsilon=\frac{h|a|}{2}$。实践也表明这些格式是可用于穿行计算的。此外，由于间断条件与微分方程一样都是从积分守恒律导出的，因此用于穿行计算的差分格式最好也直接从积分守恒律推导，也就是说，最好取为守恒型的。事实上，守恒型差分格式的引入正是为了适应间断解的计算的。间断解特别是冲击波的计算是一个重要的课题，可以参考[1]。

参 考 资 料

[1] 李奇特迈尔，《初值问题差分方法》，科学出版社，1966；或其新版，Richtmyer-Morton, "Difference Methods for Initial Value Problems," 2nd ed., 1967。

[2] 儒科夫，《应用特征线方法解气体力学一维问题》，上海科学技术出版社，1963。

[3] 张关泉，《关于气动力学方程的一个差分格式》，应用数学与计算数学 1:1(1964)，57～68 页。

第十三章 偏微分方程边值问题数值解法

§13.1 问题的来源

13.1.1 椭圆方程及其定解条件

物理上的定常态问题，例如弹性力学中的平衡问题，无粘性流体的无旋运动，亚声速流，位势场包括静电磁场、引力场等，热传导(温度分布)及扩散(浓度分布)问题等等通常归结为椭圆型微分方程。最简单的典型就是拉普拉斯方程

$$-\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=0 \tag{13.1.1}$$

这是齐次的，即不带右项。还有泊松方程

$$-\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=f \tag{13.1.2}$$

这是带有右项即非齐次的方程。更一般些则有变系数的方程如

$$-\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)=f \tag{13.1.3}$$

这里 $\beta=\beta(x, y)>0$, $f=f(x, y)$ 是给定的系数分布。

椭圆方程的主要定解问题是边值问题，即要求定出未知函数 $u=u(x, y)$ 使它在某区域 Ω 内满足微分方程如(13.1.3)，并且在边界 $\partial\Omega$ 上满足给定的边界条件。方程(13.1.3)是二阶的，需要给出一个边界条件，通常取下列三类形式之一，而且在边界的不同区段上可以取不同类的条件。

第一类：给定函数值如

$$u=\bar{u}$$

第二类：给定外法向导数值如

$$\beta\frac{\partial u}{\partial \nu}=q$$

第三类：给定函数及外法向导数的线性组合的值如

$$\beta\frac{\partial u}{\partial \nu}+\eta u=q,\quad \eta\geqslant 0$$

这里 $\bar{u}$, q, β, η 均为边界上给定的分布，β 就是方程(13.1.3)中的系数在边界上所取的值，$\beta>0$。

显然可见，第二类是第三类条件的特殊情况，相当于 $\eta\equiv 0$。此外，由于 $\beta>0$，第一类虽然不能被包括在第三类之内，但它可以看为第三类边界条件

$$\beta\frac{\partial u}{\partial \nu}+\eta(u-\bar{u})=0$$

当 $\eta\to\infty$ 的极限情况。

一般说来，边界条件可以表为

$$\partial\Omega=\Gamma_0+\Gamma_0'$$

$$\Gamma_0:\ u=\bar{u} \tag{13.1.4}$$

$$\Gamma_0':\ \beta\frac{\partial u}{\partial \nu}+\eta u=q \tag{13.1.5}$$

即 $\partial\Omega$ 分解为互补的两个部分 Γ_0, Γ_0',其上分别给定第一类和第三类边界条件。Γ_0, Γ_0' 各自又可以是由不相连的区段组成。图 13.1 中边界上组线部分表示 Γ_0, 其余的部分表示 Γ_0'。

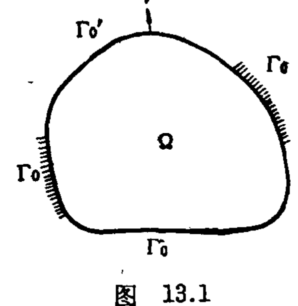


图 13.1

13.1.2 守恒原理

椭圆方程的数值解法分为两步。首先是把它离散化，变为一组代数方程。然后解算这组代数方程。

为了搞好离散化，最好要懂得待解的微分方程在物理上是怎样导来的。这样对问题背景有较深的理解后可以减少解题时的盲目性；而且，以后将要表明，比较正确的离散化方法正是从问题的推导过程中自然形成的，并且由此得到的代数方程在解算时也是比较有利的。

物理上有意义的微分方程总是某种守恒规律（包括平衡规律，协调规律等等）的数学表达形式。它们是从积分形式的守恒原理推导出来的。此外，在大多数场合下，它们还可以从另一途径，即某种“能量”极值原理即变分原理导出。下面将以方程(13.1.3)为例说明，物理有众多的现象都可以用 (13.1.3) 来描述。为了方便，这里认为它是二维定常热传导方程。$u=u(x, y)$表示温度分布，$\beta=\beta(x, y)$表示介质的传热系数。

我们将从热量守恒原理来推导(13.1.3)。

设有某种介质，占有区域 Ω, 其上有温度分布 $u=u(x, y)$。温度分布的不均匀性引起热流。按照傅立叶热传导定律，在介质的任意点上，单位时间内通过法向为 $n=(n_x, n_y)$的单位截段的热量为

$$-\beta\frac{\partial u}{\partial n}$$

即正比于温度梯度，比例常数 $\beta>0$ 为介质的传热系数，负号系表示热量是从热处流向冷处。这里设介质为各向同性的，介质可以均匀的即 β 为常数，也可以非均匀的即变系数 $\beta=\beta(x,y)$。更设介质含有热源，在单位时间单位面积释放热量 $f=f(x, y)$。

在域 Ω 内任取子域 D, 图 13.2。单位时间内通过 D 以边界 ∂D 流出 D 的热量为

$$-\oint_{\partial D}\beta\frac{\partial u}{\partial n}ds$$

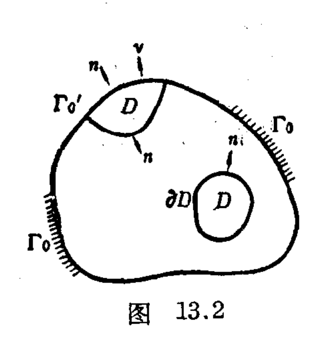


图 13.2

n 表示 ∂D 的外法向。同时间内 D 的内部热源释放的能量为

$$\iint_D f\,dx\,dy$$

由于热量的守恒性，在定常状态时应有

$$-\oint_{\partial D}\beta\frac{\partial u}{\partial n}ds=\iint_D f\,dx\,dy \tag{13.1.6}$$

这个积分关系式对于任意子域 $D\subset\Omega$ 都成立。这就是积分形式的热量守恒律。

当场量 u 充分光滑时，高斯积分公式

$$-\oint_{\partial D}\beta\frac{\partial u}{\partial n}ds=-\iint_D\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)dx\,dy$$

成立，于是(13.1.6)成为

$$\iint_D \left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial y}+f\right)dx\,dy=0 \tag{13.1.7}$$

对于一切 $D\subset\Omega$ 都成立。因此，把 D 缩至任取的点 $(x, y)\in\Omega$，“脱括弧”后即得

$$\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial y}+f=0$$

就是(13.1.3)，它在任意点 $(x, y)\in D$ 都成立。这就是微分形式的热量守恒律，通常叫做定常热传导方程。反之，设微分方程(13.1.3)在域 Ω 内成立。任取子域 $D\subset\Omega$，将(13.1.3)式两端对 D 积分，并运用高斯积分公式就能导出积分关系式(13.1.6)。因此可以说守恒律的积分形式(13.1.6)与微分形式(13.1.3)是等价的。但这种等价性是有条件的，它依赖于高斯积分公式，而后者只当场量足够光滑时才成立。在场量不够光滑处微分守恒律则取另外的形式，现以介质系数 β 有间断的情况说明之。

设 Ω 内有介质间断线，即系数 β 跨过 L 时有跳跃。为简便计，设 L 把 Ω 分割为两部 Ω^-, Ω^+，在 L 上规定指向 Ω^+ 的方向为正法向 ν 任取线段 $S\subset L$，取狭条形子域 $D\subset\Omega$ 使得 S 位于条形域 D 的中线。D 的四个顶点记为 a, b, c, d (图 13.3)，命带宽 $\overline{bc}\approx\overline{da}\approx h$。根据(13.1.6)

$$\oint_{\partial D}\beta\frac{\partial u}{\partial n}ds+\iint_D f\,dx\,dy=\oint_{ab+bc+cd+da}\beta\frac{\partial u}{\partial n}ds$$
$$+\iint_D f\,dx\,dy=0 \tag{13.1.8}$$

图 13.3

设源项 f 在 L 上没有集中的奇异性。于是当 $h\to0$ 时

$$\iint_D f\,dx\,dy\to0,\ \int_{bc+da}\beta\frac{\partial u}{\partial n}ds\to0,\ \int_{ab+cd}\beta\frac{\partial u}{\partial n}ds\to\int_S\left[\left(\beta\frac{\partial u}{\partial\nu}\right)^+-\left(\beta\frac{\partial u}{\partial\nu}\right)^-\right]ds$$

因此

$$\int_S\left[\left(\beta\frac{\partial u}{\partial\nu}\right)^+-\left(\beta\frac{\partial u}{\partial\nu}\right)^-\right]ds=0$$

对任意 $S\subset L$ 成立。再将 S 缩至一个点则得

$$\left(\beta\frac{\partial u}{\partial\nu}\right)^+=\left(\beta\frac{\partial u}{\partial\nu}\right)^- \tag{13.1.9}$$

对一切点 $(x, y)\in L$ 成立。这就是在介质间断上守恒律的微分形式。由于在 L 的两侧 $\beta^+\neq\beta^-$，故法向导数也间断 $\left(\frac{\partial u}{\partial\nu}\right)^+\neq\left(\frac{\partial u}{\partial\nu}\right)^-$，但它们的乘积 $\beta\frac{\partial u}{\partial\nu}$ 则恒连续，在热传导问题中这就是热流的连续性。此外，基于物理上的考虑，温度场即函数 u 在介质间断线 L 上应该连续即 $u^+=u^-$。

这样，当介质有间断时，微分守恒律的完整形式是

$$\Omega-L:\ -\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)=f \tag{13.1.10}$$

$$L:\ \left(\beta\frac{\partial u}{\partial\nu}\right)^--\left(\beta\frac{\partial u}{\partial\nu}\right)^+=0 \tag{13.1.11}$$

(这里 $\Omega-L$ 表示域 Ω 内除去间断线 L 以外的部分)，它们综合起来与积分守恒律(13.1.6)即

$$\oint_{\partial D}\beta\frac{\partial u}{\partial n}ds+\iint_D f\,dx\,dy=0,\ \text{对一切 } D\subset\Omega \tag{13.1.12}$$

等价。

热传导问题的边界条件大致有三种。一是给定边界温度 $u=\bar{u}$，这属于第一类。二是给定通过边界的热流 $\beta\dfrac{\partial u}{\partial\nu}=q$，属于第二类，当 $q=0$ 时相当于热绝缘。三是牛顿冷却定律

$$\beta\frac{\partial u}{\partial\nu}=\eta(u^*-u)+p$$

这里 u^* 为环境温度，$\eta\geqslant 0$ 为介质与环境的热交换系数，p 为边界上的线状热源，这属于第三类，u^*，η，p 都是预给的分布。因此可以回到在 13.1.1 节中讨论的一般情况，边界条件统一地表为

$$\partial\Omega=\Gamma_0+\Gamma_0'$$

$$\Gamma_0:\ u=\bar{u} \tag{13.1.13}$$

$$\Gamma_0':\ \beta\frac{\partial u}{\partial\nu}+\eta u=q \tag{13.1.14}$$

应该指出，第二、三类边界条件可以自然地吸收在积分守恒律(13.1.6)中。事实上，在 Ω 内任取邻接于边界 $\partial\Omega$ 的子域 D 图 13.2 上方，设 D 的边界 ∂D 与 $\partial\Omega$ 上的第三类边界条件段 Γ_0' 有公共部分，记为 $\partial D\cdot\Gamma_0'$，其余的部分记为 $\partial D-\Gamma_0'$。于是由(13.1.6)自然有

$$\oint_{\partial D}\beta\frac{\partial u}{\partial n}ds=\int_{\partial D-\Gamma_0'}\beta\frac{\partial u}{\partial n}ds+\int_{\partial D\cdot\Gamma_0'}\beta\frac{\partial u}{\partial\nu}ds=\int_{\partial D-\Gamma_0'}\beta\frac{\partial u}{\partial n}ds+\int_{\partial D\cdot\Gamma_0'}[-\eta u+q]ds$$

因此，对于这样的 $D\subset\Omega$(13.1.6)就成为

$$\int_{\partial D-\Gamma_0'}\beta\frac{\partial u}{\partial n}ds+\int_{\partial D\cdot\Gamma_0'}[-\eta u+q]\,ds+\iint_D f\,dx\,dy$$

比较(13.1.10～11)和(13.1.12)可以看到：积分守恒律与微分守恒律等价，前者取积分形式，只含一阶导数，比较统一简单，并能把通常较难处理的第二、三类边界条件和交界条件自然统一在内；而后者含二阶导数，形式比较繁琐。此外，处理积分总比微分容易，处理低阶微分总比高阶容易。因此直接从积分守恒原理出发来进行离散是有利的，见 §13.3。

13.1.3 变分原理

椭圆方程(13.1.3)连同它的第二、三类边界条件也可以从适当的“变分原理”导出。事实上，对应于(13.1.3)和(13.1.5)可以构造所谓“能量积分”

$$J(u)=\iint_\Omega\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy+\int_{\Gamma_0'}\left[\frac{1}{2}\eta u^2-qu\right]ds \tag{13.1.15}$$

对于任给函数 $u=u(x,\ y)$ 有一个积分值 $J(u)$ 与之相应，因此 $J(u)$ 可以说是“函数的函数”，通常也叫做“泛函”。重要之点在于：在所有满足边界条件(13.1.4)

$$\Gamma_0:\ u=\bar{u}$$

的函数类中，使得 J 达到极小值的函数即极值函数 $u=u(x,\ y)$ 必定在 Ω 内满足微分方程(13.1.3)，而且，除了在边界段 Γ_0 上满足给定的条件(13.1.4)外还在其余的边界段 Γ_0' 上自动满足条件(13.1.5)。反之，也可以证明，设函数 $u=u(x,\ y)$ 在 Ω 内满足方程(13.1.3)，在 Γ_0，Γ_0' 上分别满足边界条件(13.1.4)，(13.1.5)则它必定是满足条件(13.1.4)的函数类使 J 达到极小的函数。这就是说条件变分问题

$$\begin{cases} J(u)=\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dxdy+\int_{\Gamma_0'}\left[\frac{1}{2}\gamma_1 u^2-qu\right]ds=\text{极小} \\ \Gamma_0:\ u=\bar{u} \end{cases} \tag{13.1.16}$$

等价于边值问题

$$\begin{cases} \Omega:\ -\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)=f \\ \Gamma_0':\ \beta\frac{\partial u}{\partial \nu}+\gamma_1 u=q \\ \Gamma_0:\ u=\bar{u} \end{cases} \tag{13.1.17}$$

(参看[1],[4]或第十四章§14.1)

应该指出，上述等价性是在介质系数β无间断时成立的，这是因为与13.1.1节中情况相似，等价性的过渡有赖于高斯积分公式，而后者是有条件的。事实上我们知道，当系数β有间断时，在间断线L上微分方程(13.1.3)不成立而应代以交界条件(13.1.9)。可以证明，当介质系数β有间断时，变分问题等价于下列边值问题

$$\begin{cases} \Omega-L:\ -\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)=f \\ L:\ \left(\beta\frac{\partial u}{\partial \nu}\right)^{-}-\left(\beta\frac{\partial u}{\partial \nu}\right)^{+}=0 \\ \Gamma_0':\ \beta\frac{\partial u}{\partial \nu}+\gamma_1 u=q \\ \Gamma_0:\ u=\bar{u} \end{cases} \tag{13.1.18}$$

从上述等价性可以看到：在解微分方程边值问题时，第二、三类边界条件以及当有介质间断时的交界条件都必须作为定解条件列出。但是，在解等价的变分问题时，这些条件是被J的极值函数所自动满足的，无须作为定解条件列出，这就大大简化了。这类边界条件叫做自然的边界条件。与此相反，第一类边界条件，在变分问题中和在微分方程问题一样需要作为定解条件列出，这类边界条件叫做强加的边界条件。对于方程(13.1.18)，强加边界条件只含u本身，比较简单，而自然边界条件含有法向导数，一般较难处理，特别当几何形状较复杂时。能量积分(13.1.16)中只含一阶导数，比方程(13.1.18)中的二阶导数要容易处理。因此直接从变分原理出发来解题也是一条有利的途径，见§13.4。特别是第十四章有限元方法。变分原理的有利因素与守恒原理是基本相似的(比较13.1.2节之末段)。

§13.2 离散化和差分格式

椭圆方程的待定解是连续变量的未知函数，有无穷多个自由度。为了进行数值解，首先要把问题离散化，把无穷多自由度的问题简化为有限多个自由度的问题，也就是把微分方程代为一组代数方程，然后进行解算。

离散化的一类方法是差分方法。问题的定解区域化为离散的格网，待定函数代为格网的节点值，微分方程和边界条件等则化为相应的差分方程。这类方法简单通用，在进入计算机的时代后成为主要的数值方法。

离散化的另一类方法是函数的有限展开方法，亦称李兹-加辽金方法。适当选取一组基函数$\varphi_1, \varphi_2, \cdots$，把待定解展为有限的线性组合$u=\Sigma a_i\varphi_i$，要求在一定的意义下近似满足微

分方程和定解条件，或近似满足相应的变分原理，从而得出关于系数 a_1, a_2, …的代数方程组以定解。这类方法在历史上曾占重要地位。但是，在其传统的形式下，基函数的选择往往很困难，通用性较差，故对此将不作专门介绍。

近年来还发展了一类离散化方法即有限元法。这是以格网为基础的基函数展开法。可以说它是前述两类方法的比较成功的结合，在椭圆型问题已上升到主导地位，特别适合于几何上物理上比较复杂的问题，也便于算法的标准化，将在第十四章中介绍。也是由于这个原因，本章对于差分方法的讨论主要局限于比较简单规则的问题。

求解矩形域 $\Omega=[a, b, c, d]$ 上的泊松方程

$$-\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=f(x, y) \tag{13.2.1}$$

为了说明问题，设边界条件同含一、二两类，

$$ab, ac: \quad u=\bar{u} \tag{13.2.2}$$

$$cd, bd: \quad \frac{\partial u}{\partial \nu}=q \tag{13.2.3}$$

ν 为外法线，f, $\bar{u}$, q 为给定的分布。

最简单的离散化方法是差分化，即布上格网，把方程和边界条件中的微商代以差商而得到差分方程组。为此目的，取顶点 a 为坐标轴的原点，用纵横线

$$x=x_i=ih, \quad i=0, 1, \cdots, M$$

$$y=y_j=jk, \quad j=0, 1, \cdots, N$$

把 Ω 分为等距格网如图 13.4。h, k 分别是 x, y 方向的步长，节点 (x_i, y_j) 记为 (i, j)，函数值 $u(x_i, y_j)$ 记为 u_{ij}。节点分为两类，一类是内点 $(i=1, \cdots, M-1; j=1, \cdots, N-1)$，其余是边点。边点中有四个角点即 a, b, c, d，其它是非角点的边点。

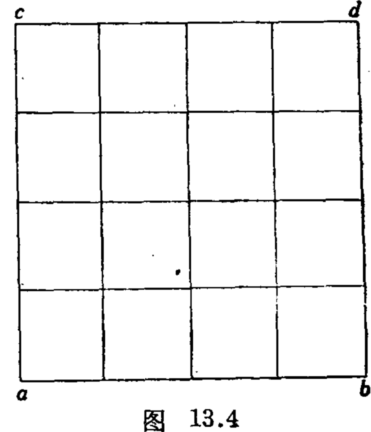
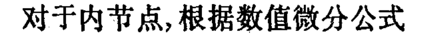


图 13.4



对于内节点，根据数值微分公式

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j}=\frac{1}{h^2}(u_{i-1,j}-2u_{i,j}+u_{i+1,j})+O(h^2)$$

$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j}=\frac{1}{k^2}(u_{i,j-1}-2u_{i,j}+u_{i,j+1})+O(k^2)$$

可将(13.2.1)代为差分方程

$$-\frac{1}{h^2}(u_{i-1,j}-2u_{i,j}+u_{i+1,j})-\frac{1}{k^2}(u_{i,j-1}-2u_{i,j}+u_{i,j+1})=f_{i,j} \tag{13.2.4}$$

即

$$2\left(\frac{1}{h^2}+\frac{1}{k^2}\right)u_{i,j}-\frac{1}{h^2}(u_{i-1,j}+u_{i+1,j})-\frac{1}{k^2}(u_{i,j-1}+u_{i,j+1})=f_{i,j} \tag{13.2.5}$$

$$i=1, \cdots, M-1; \quad j=1, \cdots, N-1$$

它联系上下左右共五个节点，叫做五点差分格式。在 $h=k$ 的情况下则简化为

$$4u_{i,j}-(u_{i-1,j}+u_{i+1,j}+u_{i,j-1}+u_{i,j+1})=h^2 f_{i,j} \tag{13.2.6}$$

至于边点，在 ab, ac 上第一类边界条件(13.2.2)可以表为

$$ab:\ u_{i,0}=\bar{v}_{i,0},\ i=0,\ 1,\ \cdots,\ M \tag{13.2.7}$$

$$ac:\ u_{0,j}=\bar{v}_{0,j},\ j=0,\ 1,\ \cdots,\ N \tag{13.2.8}$$

利用差商公式

$$\left(\frac{\partial u}{\partial x}\right)_{i,j}=\frac{1}{h}(u_{i,j}-u_{i-1,j})+O(h)$$

$$\left(\frac{\partial u}{\partial y}\right)_{i,j}=\frac{1}{k}(u_{i,j}-u_{i,j-1})+O(k)$$

可把 bd, cd 上第二类边界条件(13.2.3)表为

$$bd:\ \frac{1}{h}(u_{M,j}-u_{M-1,j})=q_{M,j},\ j=1,\ \cdots,\ N-1 \tag{13.2.9}$$

$$cd:\ \frac{1}{k}(u_{i,N}-u_{i,N-1})=q_{i,N},\ i=1,\ \cdots,\ M-1 \tag{13.2.10}$$

它联系上下或左右两点,是两点差分格式。在 $h=k$ 的情况下也可写作

$$u_{M,j}-u_{M-1,j}=hq_{M,j},\ j=1,\ \cdots,\ N-1 \tag{13.2.11}$$

$$u_{i,N}-u_{i,N-1}=hq_{i,N},\ i=1,\ \cdots,\ M-1 \tag{13.2.12}$$

还有左上方的角点 $d=(M,\ N)$。对此按 bd 或 cd 的方式处理有矛盾。但是，由于 $u_{M,N}$ 在前列其它方程中都不出现，故不妨略去这个节点。于是共有 $(M+1)(N+1)-1$ 个方程和相同数目的未知数 u_{ij} ($i=0,\ 1,\ \cdots,\ M$; $j=0,\ 1,\ \cdots,\ N$, 除去 $i=M$, $j=N$) 可以定解，具体解法见 § 13.5。

如果将四边边界条件都改为第一类

$$\partial\Omega:\ u=g \tag{13.2.13}$$

并取 Ω 为正方形,格网亦取正方的(即 $h=k$, $M=N$)。得到差分方程组

$$4u_{i,j}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}=h^2f_{i,j},\ i,\ j=1,\ \cdots,\ N-1 \tag{13.2.14}$$

$$u_{i,j}=g_{i,j},\ \text{当}\ i=0,\ N\ \text{或}\ j=0,\ N \tag{13.2.15}$$

这里可以视内点的 $u_{i,j}(i,\ j=1,\ \cdots,\ N-1)$ 为未知数，共$(N-1)^2$个，也有同数量的内点方程。可以称为模型的椭圆差分方程组。

如上所列的差分方程通常也叫做差分格式。这是因为这种公式是规格化的，人们无须把各节点的方程逐个列出，只须列少数几种典型的格式，而在同类型的节点上重复套用，这在编制程序时比较方便并有节约存储的优点。

§ 13.3 基于守恒原理的差分格式

可以直接从守恒原理 (13.1.2 节) 来推导问题 (13.2.1-3) 的差分格式。对应于方程(13.2.1)有积分关系式(13.1.6)即

$$-\oint_{\partial D}\frac{\partial u}{\partial n}\,ds=\iint_D f\,dx\,dy,\ \text{对一切}\ D\subset\Omega\ \text{均成立} \tag{13.3.1}$$

在图 13.4 的格网上引进“半线”(图 13.5 中的虚线)

$$x=x_{i+\frac{1}{2}}=\left(i+\frac{1}{2}\right)h,\ i=0,\ \cdots,\ M-1$$

$$y=y_{j+\frac{1}{2}}=\left(j+\frac{1}{2}\right)k,\ j=0,\ \cdots,\ N-1$$

这些半线形成另一套格网，与原格网相辅相成。对应于每个节点(i, j)有一个虚矩形$D_{i,j}$，如图13.5中的阴影块。注意它们在边点上只有半块大，在角点上只有四分之一块大。

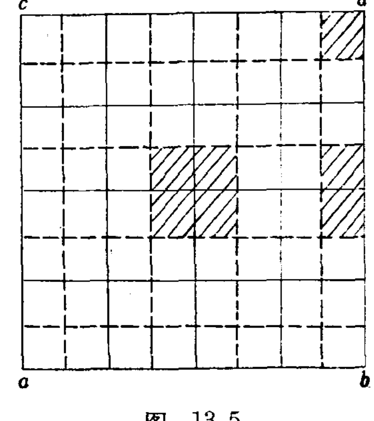


图 13.5

$D_{i,j}(i=0, \cdots, M; j=0, \cdots, N)$形成$\Omega$的无遗漏、无重复、无多余的覆盖。作为离散化的第一步，要求积分关系式对于每个$D_{i,j}$成立。

$$-\oint_{\partial D_{i,j}} \frac{\partial u}{\partial n} ds = \iint_{D_{i,j}} f\, dx\, dy, \qquad (13.3.2)$$

$$i=0, \cdots, M; \quad j=0, \cdots, N$$

为简便计只讨论$h=k$的情况，对$h\neq k$乃至非均匀步长的情况本质相同，只是系数复杂一些。

命

$$A_{i+\frac{1}{2},j} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\frac{\partial u}{\partial x}\right)_{x=x_{i+\frac{1}{2}}} dy \approx \frac{k}{h}(u_{i+1,j}-u_{i,j}) = u_{i+1,j}-u_{i,j}$$

$$B_{i,j+\frac{1}{2}} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\frac{\partial u}{\partial y}\right)_{y=y_{j+\frac{1}{2}}} dx \approx \frac{h}{k}(u_{i,j+1}-u_{i,j}) = u_{i,j+1}-u_{i,j}$$

$$F_{i,j} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f\, dx\, dy \approx hkf_{i,j} = h^2 f_{i,j}$$

于是对于内点(i, j)，(13.3.2)可以表为(见图13.5的中央)

$$A_{i-\frac{1}{2},j} - A_{i+\frac{1}{2},j} + B_{i,j-\frac{1}{2}} - B_{i,j+\frac{1}{2}} = F_{i,j} \qquad (13.3.3)$$

即

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = h^2 f_{i,j} \qquad (13.3.4)$$

对于右边点(M, j)，$j=1, \cdots, N-1$　命

$$A_{M,j} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\frac{\partial u}{\partial x}\right)_{x=x_M} dy = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\frac{\partial u}{\partial \nu}\right)_{x=x_M} \approx kq_{M,j} = hq_{M,j}$$

$$B_{M,j+\frac{1}{2}} = \int_{x_{M-\frac{1}{2}}}^{x_M} \left(\frac{\partial u}{\partial y}\right)_{y=y_{j+\frac{1}{2}}} dx \approx \frac{h}{2k}(u_{M,j+1}-u_{M,j}) = \frac{1}{2}(u_{M,j+1}-u_{M,j})$$

$$F_{M,j} = \int_{x_{M-\frac{1}{2}}}^{x_M} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f\, dx\, dy \approx \frac{1}{2} hkf_{M,j} = \frac{1}{2} h^2 f_{M,j}$$

这里自然地吸收了第二类边界条件，于是(13.3.2)表为(见图13.5的右端)

$$A_{M-\frac{1}{2},j} - A_{M,j} + B_{M,j-\frac{1}{2}} - B_{M,j+\frac{1}{2}} = F_{M,j} \qquad (13.3.5)$$

即

$$2u_{M,j} - u_{M-1,j} - \frac{1}{2} u_{M,j-\frac{1}{2}} - \frac{1}{2} u_{M,j+\frac{1}{2}} = \frac{1}{2} h^2 f_{M,j} + hq_{M,j}, \quad j=1, \cdots, N-1 \qquad (13.3.6)$$

对于上边点(i, N)，$i=1, \cdots, M-1$，情况也类似，命

$$A_{i+\frac{1}{2},N} = \int_{y_{N-\frac{1}{2}}}^{y_N} \left(\frac{\partial u}{\partial x}\right)_{x=x_{i+\frac{1}{2}}} dy \approx \frac{k}{2h}(u_{i+1,N}-u_{i,N}) = \frac{1}{2}(u_{i+1,N}-u_{i,N})$$

$$B_{i,N} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\frac{\partial u}{\partial y}\right)_{y=y_N} dx = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\frac{\partial u}{\partial \nu}\right)_{y=y_N} dx \approx hq_{i,N}$$

$$F_{i,N}=\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{N-\frac{1}{2}}}^{y_N} f\,dx\,dy\approx\frac{1}{2}hkf_{i,N}=\frac{1}{2}h^2f_{i,N}$$

则(13.3.2)表为

$$A_{i-\frac{1}{2},N}-A_{i+\frac{1}{2},N}+B_{i,N-\frac{1}{2}}-B_{i,N}=F_{i,N} \tag{13.3.7}$$

即

$$2u_{i,N}-\frac{1}{2}u_{i-1,N}-\frac{1}{2}u_{i+1,N}-u_{i,N-1}=\frac{1}{2}h^2f_{i,N}+hq_{i,N},\quad j=1,\ \cdots,\ N-1 \tag{13.3.8}$$

对于角点$(M,\ N)$命

$$A_{M,N}=\int_{y_{N-\frac{1}{2}}}^{y_N}\left(\frac{\partial u}{\partial x}\right)_{x=x_M}dy=\int_{y_{N-\frac{1}{2}}}^{y_N}\left(\frac{\partial u}{\partial \nu}\right)_{x=x_M}dy\approx\frac{1}{2}kq_{M,N}=\frac{1}{2}hq_{M,N}$$

$$B_{M,N}=\int_{x_{M-\frac{1}{2}}}^{x_M}\left(\frac{\partial u}{\partial y}\right)_{y=y_N}dx=\int_{x_{M-\frac{1}{2}}}^{x_M}\left(\frac{\partial u}{\partial \nu}\right)_{y=y_N}dx\approx\frac{1}{2}hq_{M,N}$$

$$F_{M,N}=\int_{x_{M-\frac{1}{2}}}^{x_M}\int_{y_{N-\frac{1}{2}}}^{y_N} f\,dx\,dy\approx\frac{1}{4}hkf_{M,N}=\frac{1}{4}h^2f_{M,N}$$

这里，为了简便，假定在角点给的 x, y 方向两个法向导数取相同值 $q_{M,N}$。于是 (13.3.2) 表为(见图 13.5 的右上角)

$$A_{M-\frac{1}{2},N}-A_{M,N}+B_{M,N-\frac{1}{2}}-B_{M,N}=F_{M,N} \tag{13.3.9}$$

即

$$u_{M,N}-\frac{1}{2}u_{M-1,N}-\frac{1}{2}u_{M,N-1}=\frac{1}{4}h^2f_{M,N}+\frac{1}{2}hq_{M,N}+\frac{1}{2}hq_{M,N} \tag{13.3.10}$$

至于左边及下边第一类边界条件则同(13.2.7～13.2.8)。

以上 (13.3.3～13.3.10) 每个方程都是反映了单元 $D_{i,j}$ 的局部守恒性。由于交界项 $A_{i+\frac{1}{2},j}$ 在相邻单元 $D_{i,j}$ 及 $D_{i+1,j}$ 的方程中反号，$B_{i,j+\frac{1}{2}}$ 在相邻的 $D_{i,j}$ 及 $D_{i,j+1}$ 的方程中反号，因此将相邻单元的方程相加时，其交界项正好抵消，这意味着甲方“给予”乙方的热量等于乙方从甲方“收到”的热量，这种“收支”平衡的关系正是守恒性的要点。因此局部的守恒性保证了整体的守恒性。例如，对于 $i=i',\ i'+1,\ \cdots,\ i''$；$j=j',\ j'+1,\ \cdots,\ j''$ 的方程相累加，则所有内部交界项都抵消而得到

$$\sum_{j=j'}^{j''}A_{i'-\frac{1}{2},j}-\sum_{j=j'}^{j''}A_{i''+\frac{1}{2},j}+\sum_{i=i'}^{i''}B_{i,j'-\frac{1}{2}}-\sum_{i=i'}^{i''}B_{i,j''+\frac{1}{2}}=\sum_{i=i'}^{i''}\sum_{j=j'}^{j''}F_{i,j} \tag{13.3.11}$$

这相当于区域

$$D=\sum_{i=i'}^{i''}\sum_{j=j'}^{j''}D_{i,j} \tag{13.3.12}$$

上的守恒关系式(13.3.1)。

利用(13.2.7-8)代进第一类边界点的已知值，得到 MN 个未知数 $u_{11}, u_{12}, \cdots, u_{MN}$ 和相同数目的方程(13.3.4-6-8-10)，可以定解。如果将未知数依序记为 $u_1,\ u_2,\ \cdots,\ u_n,\ n=MN$，将各方程也按同一次序排列并将已知项统统移到右端则得

$$\sum_{j=1}^{n}a_{ij}u_j=b_i,\quad i=1,\ \cdots,\ n \tag{13.3.13}$$

即

$$\boldsymbol{Au}=\boldsymbol{b} \tag{13.3.14}$$

可以证明,系数阵 A 是对称

$$a_{ij}=a_{ji}, i, j=1, \cdots, n$$

正定的(见§13.4),因此(13.3.13)有唯一解。这里的对称性是显然的，它意味着：当节点 i 与 j 不相邻时,点 i 方程中点 j 的系数和点 j 方程中点 i 的系数同为0。当节点 i 与 j 相邻时,则点 i 方程中点 j 的系数等于点 j 方程中点 i 的系数,这可以从方程(13.3.3~10)中看出来,实际上反映了差分方程的守恒型即甲方给予乙方的等于乙方受之于甲方的。

把这种基于守恒原理的差分格式与§13.2中基于数值微分的格式相比较，在内点是相同的，不同之处只在第二类边点。在§13.2中第二类边界条件(13.2.3)是与基本方程(13.2.1)割裂开来的形式化的处理，而在这里是与基本守恒律统一起来考虑的，比较优越。象守恒这样的基本规律性在离散化时能得到保持,是很重要的。

守恒方法显然可以推广到变系数情况。这时引用积分关系式(13.3.2)而取

$$A_{i+\frac{1}{2},j}\approx\beta_{i+\frac{1}{2},j}(u_{i+1,j}-u_{i,j})$$

$$B_{i,j+\frac{1}{2}}\approx\beta_{i,j+\frac{1}{2}}(u_{i,j+1}-u_{i,j})$$

等等,无待细说。

守恒方法处理不规则边界上的第二、三类边界条件也是比较方便的,从图13.6就可以看清楚。对于这种情况,单纯用数值微分的方法是难以处理得当的。当问题趋于复杂时,守恒方法的优点就更显著。

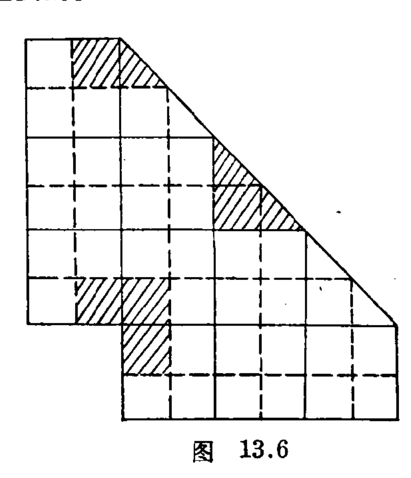


图 13.6

关于第二类边值问题

考虑问题(13.2.1-3)的一个变形,即全部边界条件都是第二类:

$$\Omega:\quad -\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=f \tag{13.3.15}$$

$$\partial\Omega:\quad \frac{\partial u}{\partial \nu}=g \tag{13.3.16}$$

通常叫做第二类边值问题或纽曼(Neuman)问题。这时问题成为退化的，可以没有解，有解时也不唯一:

(1) 相应的齐次问题——即 $f\equiv0$, $g\equiv0$ 的情况——显然有非零解

$$u\equiv c=\text{常数} \tag{13.3.17}$$

并且可以证明齐次问题的解必定作此形式。

(2) 问题并不是对随便的 f, g 都有解。事实上对(13.3.15)两端在 Ω 上积分，运用高斯公式,得到

$$-\oint_{\partial\Omega}\frac{\partial u}{\partial \nu}ds=\iint_{\Omega}f\,dx\,dy$$

这相当于在(13.3.1)中取 $D=\Omega$,再用条件(13.3.16)就得到

$$\iint_{\Omega}f\,dx\,dy+\oint_{\partial\Omega}g\,ds=0 \tag{13.3.18}$$

这就是说，当问题有解时，f 和 g 必须满足这一条件。反之，可以证明，当条件(13.3.18)满足时,问题(13.3.15~16)有解;命 $\tilde{u}$ 为一个特解,则问题的通解必可表为特解 $\tilde{u}$ 与相应齐次

问题的通解之和，即

$$u=\tilde{u}+c \tag{13.3.19}$$

条件(13.3.18)通常叫做边值问题(13.3.15~16)的协调条件。它在物理上是显然的，如果将方程(13.3.15)理解为不受支承的薄膜平衡问题，f 与 g 分别是面载荷与边线载荷，显然只有当外部载荷 f 与 g 达成平衡即满足(13.3.18)时，薄膜才能达成平衡。(13.3.17)则表示任何刚性位移是不受支承不受载荷的自由薄膜的平衡解。

用守恒原理的方法离散化后得到类似于(13.3.3~10)的差分方程组，但有 $(M+1)(N+1)$ 个未知数 $u_{00}, u_{01}, \cdots, u_{MN}$，方程的个数也相同。将未知数依次排列记为 $u_1, \cdots, u_n$ 各方程也相应排列并移项后得

$$\sum_{j=1}^{n} a_{ij}u_j=b_i,\quad i=1,\ \cdots,\ n,\quad n=(M+1)(N+1) \tag{13.3.20}$$

即

$$\boldsymbol{Au}=\boldsymbol{b} \tag{13.3.21}$$

这时方程组与(13.3.13)不同之点在于退化，可以没有解，有解时也不唯一。

(3) 相应的齐次代数方程组——即 $\boldsymbol{b}=0$ 的情况——显然有非零解(从方程(13.3.3~10)的具体形式看出)

$$u_i=c=\text{常数},\quad i=1,\ \cdots,\ n \tag{13.3.22}$$

并可证明，齐次方程组的解必作如此形式。可以证明 $\boldsymbol{A}$ 为对称，退化半正定。所谓退化是指系数行列式为零

$$|\boldsymbol{A}|=0$$

(4) 方程组(13.3.21)不是对随便的右项 $\boldsymbol{b}$ 都有解。事实上，将全部方程(13.3.21)累加；这就是(13.3.11~12)而取 $i'=0$, $i''=M$, $j'=0$, $j''=N$, $D=\Omega$，并将已知项(涉及 f 和 g 的项)移至右端，由于守恒性，内部交界项都抵消而得

$$\sum_{i=1}^{n} b_i=0 \tag{13.3.23}$$

这就是说，当方程组(13.3.21)有解时，右项 b 所必需满足的条件。反之也不难证明，当(13.3.23)满足时，方程组(13.3.21)必有解；命 $\tilde{u}$ 为一个特解，则通解 $\boldsymbol{u}$ 必可表为特解 $\boldsymbol{u}$ 与齐次方程组的通解之和即

$$u_i=\tilde{u}_i+c,\quad i=1,\ \cdots,\ n \tag{13.3.24}$$

条件(13.3.23)称为退化差分方程组(13.3.21)的协调条件这一条件也可以从线代数方程组的一般理论直接导出：对称退化线性方程组(13.3.21)有解的充要条件是右项向量 $\boldsymbol{b}=(b_1, \cdots, b_n)$ 与相应齐次方程组的一切解 $\boldsymbol{v}=(v_1, \cdots, v_n)$——现在是(13.3.24)即 $\boldsymbol{v}=(c, c, \cdots, c)$——相正交，即

$$\sum_{i=1}^{n} b_i v_i=\sum_{i=1}^{n} b_i c=0$$

这就是(13.3.23)。

注意第(1)，(2)段的(13.3.17)，(13.3.18)，(13.3.19)分别对应于第(3)，(4)段的(13.3.22)，(13.3.23)，(13.3.24)。特别从离散化的过程可以看出

$$\sum_{i=1}^{n} b_i\sim\iint_{\Omega} f\,dx\,dy+\oint_{\partial\Omega} g\,ds$$

因此对于第二类边值问题而言，应该注意到使得协调条件(13.3.18)在离散的意义下也得到满足，即(13.3.23)应该成立以保证退化线代数方程组有解。如果某种离散化处理所得右项 b_i 不满足 (13.3.23) 即 $\sum_{i=1}^{n} b_i=\varepsilon\neq 0$，则可以用 $b_i-\frac{\varepsilon}{n}$ 代替 b_i 作为新的右项而满足协调条件(13.3.23)从而保证方程组可解。

§13.4 基于变分原理的差分格式

可以从变分原理出发来形成差分格式。根据13.1.3节边值问题(13.2.1-3)等价于条件变分问题

$$\begin{cases} J(u)=\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy-\int_{\Gamma_0'} qu\,ds=\text{极小} \\ \Gamma_0:\quad u=\bar{u} \end{cases} \tag{13.4.1}$$

这里 Γ_0 就是边线 $ab+ac$，Γ_0' 就是边线 $bd+cd$。

取格网如图13.4。纵横线 $x=x_i=ih$，$y=y_j=jk$，把 Ω 剖分为矩形块即面单元

$$C_{i+\frac{1}{2},j+\frac{1}{2}}=\{x_i\leqslant x\leqslant x_{i+1},y_j\leqslant y\leqslant y_{j+1}\}$$

$$\Omega=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}C_{i+\frac{1}{2},j+\frac{1}{2}}$$

同时也把边线 Γ_0' 剖分为线段即线单元

$$C_{M,j+\frac{1}{2}}=\{x=x_M,\ y_j\leqslant y\leqslant y_{j+1}\}$$

$$C_{i+\frac{1}{2},N}=\{x_i\leqslant x\leqslant x_{i+1},\ y=y_N\}$$

$$\Gamma_0'=\sum_{j=0}^{N-1}C_{M,j+\frac{1}{2}}+\sum_{i=0}^{M-1}C_{i+\frac{1}{2},N}$$

于是

$$J(u)=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}J_{i+\frac{1}{2},j+\frac{1}{2}}+\sum_{j=0}^{N-1}J_{M,j+\frac{1}{2}}+\sum_{i=0}^{M-1}J_{i+\frac{1}{2},N} \tag{13.4.2}$$

此处

$$J_{i+\frac{1}{2},j+\frac{1}{2}}=\iint\limits_{C_{i+\frac{1}{2},j+\frac{1}{2}}}\left\{\frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy$$

$$J_{M,j+\frac{1}{2}}=-\int\limits_{C_{M,j+\frac{1}{2}}}qu\,ds$$

$$J_{i+\frac{1}{2},N}=-\int\limits_{C_{i+\frac{1}{2},N}}qu\,ds$$

进一步要对各单元的能量积分进行离散化。对此有许多途径。比方说，可以把积分号下的函数用差商的平均值代替：

$$\iint\limits_{C_{i+\frac{1}{2},j+\frac{1}{2}}}\frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2dx\,dy\approx\frac{1}{2}\left[\frac{1}{2}\left(\frac{u_{i+1,j}-u_{i,j}}{h}\right)^2+\frac{1}{2}\left(\frac{u_{i+1,j+1}-u_{i+1,j+1}}{h}\right)^2\right]hk$$

$$=\frac{1}{4}(u_{i+1,j}-u_{i,j})^2+\frac{1}{4}(u_{i+1,j+1}-u_{i,j+1})^2$$

这里，为了简便，与§13.2，§13.3一样，考虑$h=k$的情况。类似地处理关于$\frac{1}{2}\left(\frac{\partial u}{\partial y}\right)^2$的面积分。对于$fu$的面积分则可以采用，比方说，梯形公式

$$\iint\limits_{C_{i+\frac{1}{2},j+\frac{1}{2}}} fu\,dx\,dy=\frac{h^2}{4}(f_{i,j}u_{i,j}+f_{i+1,j}u_{i+1,j}+f_{i,j+1}u_{i,j+1}+f_{i+1,j+1}u_{i+1,j+1})$$

因此有

$$J_{i+\frac{1}{2},j+\frac{1}{2}}=\frac{1}{4}\Big[(u_{i+1}-u_{i,j})^2+(u_{i+1,j+1}-u_{i,j+1})^2+(u_{i,j+1}-u_{i,j})^2+(u_{i+1,j+1}-u_{i+1,j})^2\Big]$$
$$-\frac{h^2}{4}(f_{i,j}u_{i,j}+f_{i+1,j}u_{i+1,j}+f_{i,j+1}u_{i,j+1}+f_{i+1,j+1}u_{i+1,j+1}) \qquad (13.4.3)$$

类似地对qu的线积分也采用梯形公式，得到

$$J_{M,j+\frac{1}{2}}=-\frac{h}{2}(q_{M,j}u_{M,j}+q_{M,j+1}u_{M,j+1}) \qquad (13.4.4)$$

$$J_{i+\frac{1}{2},N}=-\frac{h}{2}(q_{i,N}u_{i,N}+q_{i+1,N}u_{i+1,N}) \qquad (13.4.5)$$

以上诸式中用到的$u_{0,j}$及$u_{i,0}$则根据强加边界条件 (13.2.2) 取已知值 (13.2.7～8)。这样，能量积分$J(u)$就离散化为MN个未知数$u_{i,j}(i=1,\cdots,M;j=1,\cdots,N)$的二次函数$J(u_{11},u_{12},\cdots,u_{MN})$。因此变分原理(13.4.1)就化为多元函数的极小问题

$$J(u_{11},u_{12},\cdots,u_{MN})=\text{极小} \qquad (13.4.6)$$

其极值方程为

$$\frac{\partial J}{\partial u_{ij}}=0,\ i=1,\cdots,M;\ j=1,\cdots,N \qquad (13.4.7)$$

二次函数微分后成为一次，故这是MN个线代数方程。

当(i,j)为内点，即$i=1,\cdots,M-1;j=1,\cdots,N-1$时，$J$中只有与点$(i,j)$邻接的四个面元项才含有$u_{i,j}$，

$$\begin{aligned}\frac{\partial J}{\partial u_{i,j}}&=\frac{\partial}{\partial u_{i,j}}\left\{J_{i-\frac{1}{2},j-\frac{1}{2}}+J_{i-\frac{1}{2},j+\frac{1}{2}}+J_{i+\frac{1}{2},j-\frac{1}{2}}+J_{i+\frac{1}{2},j+\frac{1}{2}}\right\}\\
&=\frac{1}{2}(u_{i,j}-u_{i-1,j})+\frac{1}{2}(u_{i,j}-u_{i,j-1})-\frac{h^2}{4}f_{i,j}\\
&\quad+\frac{1}{2}(u_{i,j}-u_{i-1,j})+\frac{1}{2}(u_{i,j}-u_{i,j+1})-\frac{h^2}{4}f_{i,j}\\
&\quad+\frac{1}{2}(u_{i,j}-u_{i+1,j})+\frac{1}{2}(u_{i,j}-u_{i,j-1})-\frac{h^2}{4}f_{i,j}\\
&\quad+\frac{1}{2}(u_{i,j}-u_{i+1,j})+\frac{1}{2}(u_{i,j}-u_{i,j+1})-\frac{h^2}{4}f_{i,j}\\
&=0\end{aligned}$$

由此得方程同于(13.3.3)。

当(i,j)为右边点而非角点即$i=M,j=1,\cdots,N-1$时，J中只有与点(M,j)邻接的两个面元项和两个线元项才含有$u_{M,j}$，

$$\frac{\partial J}{\partial u_{M,j}}=\frac{\partial}{\partial u_{M,j}}\{J_{M-\frac{1}{2},j-\frac{1}{2}}+J_{M-\frac{1}{2},j+\frac{1}{2}}+J_{M,j-\frac{1}{2}}+J_{M,j+\frac{1}{2}}\}$$
$$=\frac{1}{2}(u_{M,j}-u_{M-1,j})+\frac{1}{2}(u_{M,j}-u_{M,j-1})-\frac{h^2}{4}f_{M,j}$$
$$+\frac{1}{2}(u_{M,j}-u_{M-1,j})+\frac{1}{2}(u_{M,j}-u_{M,j+1})-\frac{h^2}{4}f_{M,j}-\frac{h}{2}q_{M,j}-\frac{h}{2}q_{M,j}$$
$$=0$$

由此得方程同于 (13.3.4)。类似地对上边的非角点 (i, N), $i=1, \cdots, M-1$ 得方程同于 (13.3.5)。

当 (i, j) 为角点 (M, N) 时, J 中只与 (M, N) 邻接的一个面元项和两个线元项才含 $u_{M,N}$,

$$\frac{\partial J}{\partial u_{M,N}}=\frac{\partial}{\partial u_{M,N}}\{J_{M-\frac{1}{2},N-\frac{1}{2}}+J_{M,N-\frac{1}{2}}+J_{M-\frac{1}{2},N}\}$$
$$=\frac{1}{2}(u_{M,N}-u_{M-1,N})+\frac{1}{2}(u_{M,N}-u_{M,N-1})-\frac{h^2}{4}f_{M,N}-\frac{h}{2}q_{M,N}-\frac{h}{2}q_{M,N}$$
$$=0$$

由此得方程同于(13.3.6)。

这样，从变分原理出发，对能量积分作适当的离散化后，可以导出一组差分格式，与在§13.3中从守恒原理导出的相同。对能量积分可以有不同的离散方案，导致大同小异的差分格式。

将未知数 $u_{i,j}(i=1, \cdots, M, j=1, \cdots, N)$ 按一定顺序排列后改记为 $u_1, u_2, \cdots, u_n$ $(n=MN)$，与此同时将(13.4.7)内的方程也按同一顺序排列，并将常数项移到右边，得到方程组

$$\sum_{j=1}^{n}a_{ij}u_j=b_i,\quad i=1, \cdots, n \tag{13.4.8}$$

或表为矩阵形式

$$\boldsymbol{Au}=\boldsymbol{b} \tag{13.4.9}$$

可以证明系数矩阵 $\boldsymbol{A}=[a_{ij}]$ 为对称正定。

事实上，能量 J，作为 $u_1, \cdots, u_n$ 的二次函数，一定能表为

$$J(u_1, \cdots, u_n)=\frac{1}{2}\sum_{i,j=1}^{n}\tilde{a}_{ij}u_iu_j-\sum_{i=1}^{n}\tilde{b}_iu_i+\tilde{c} \tag{13.4.10}$$

而(13.4.7)按顺序就是

$$\frac{\partial J}{\partial u_i}=0,\quad i=1, \cdots, n \tag{13.4.11}$$

不难算出

$$\frac{\partial J}{\partial u_i}=\frac{1}{2}\sum_{j=1}^{n}(\tilde{a}_{ij}+\tilde{a}_{ji})u_j-\tilde{b}_i$$

(13.4.11)就成为

$$\sum_{j=1}^{n}\frac{1}{2}(\tilde{a}_{ij}+\tilde{a}_{ji})u_j=b_i$$

与(13.4.8)比较系数，就得到

$$a_{ij}=\frac{1}{2}(\tilde{a}_{ij}+\tilde{a}_{ji})=a_{ji},\ b_i=\tilde{b}_i$$

因此 A 对称。由于

$$\sum_{i,j=1}^{n}\tilde{a}_{ij}u_iu_j\equiv\sum_{i,j=1}^{n}\frac{1}{2}(\tilde{a}_{ij}+\tilde{a}_{ji})u_iu_j\equiv\sum_{i,j=1}^{n}a_{ij}u_iu_j$$

所以也知道

$$J(u_1,\ \cdots,\ u_n)=\frac{1}{2}\sum_{i,j=1}^{n}a_{ij}u_iu_j-\sum_{i=1}^{n}b_iu_i+\tilde{c}$$

至于 A 的正定性，只须证明二次型

$$J_0(u_1,\ \cdots,\ u_n)=\sum_{i,j=1}^{n}a_{ij}u_iu_j$$

恒 $\geqslant 0$ 而且当 $J_0(u_1,\ \cdots,\ u_n)=0$ 时必有 $u_1=\cdots=u_n=0$。事实上，从能量函数 J 的构成公式可以知道它是以定解条件的数据 $f_{i,j},\ q_{M,j},\ q_{i,N},\ \bar{u}_{i,0},\ \bar{u}_{0,j}$ 为参数，以 $u_1,\ \cdots,\ u_n$ 为变量的二次函数。当上列参数全取 0 值时的能量函数就是 $J_0(u_1,\ \cdots,\ u_n)$，因此

$$J_0(u_1,\ \cdots,\ u_n)=\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}J_{i+\frac{1}{2},j+\frac{1}{2}}$$

其中涉及的 $f_{i,j}$ 以及 $u_{i,0}$ 和 $u_{0,j}$ 均取为 0。根据(13.4.3)，$J_{i+\frac{1}{2},j+\frac{1}{2}}$ 为平方和，因此 J_0 恒 $\geqslant 0$。此外，当 $J_0=0$ 时 $J_{i+\frac{1}{2},j+\frac{1}{2}}$ 全为 0，因此

$$u_{i,j}=u_{i+1,j}=u_{i,j+1}=u_{i+1,j+1},\quad i=0,\ \cdots,\ M-1;\ j=0,\ \cdots,\ N-1$$

由于 $u_{i,0},u_{0,j}$ 全为 0，所以 $u_{i,j}\equiv 0$ 即 $u_1=\cdots=u_n=0$。

设将问题(13.2.1～3)改为第二类边值问题(13.3.15～16)，则后者等价于无条件变分问题

$$J(u)=\iint_{\Omega}\left\{\frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy-\oint_{\partial\Omega}g\,ds=\min \qquad (13.4.12)$$

而不受任何强加约束。用变分方法离散化时，未知数有 $(M+1)(N+1)$ 个 $u_{00},\ u_{01},\ \cdots u_{MN}$，记为 $u_1,\ \cdots,\ u_n$。于是同样有

$$J(u_1,\ \cdots,\ u_n)=\frac{1}{2}\sum_{i,j=1}^{n}a_{ij}u_iu_j-\sum_{i=1}^{n}b_iu_i$$

$$J_0(u_1,\ \cdots,\ u_n)=\frac{1}{2}\sum_{i,j=1}^{n}a_{ij}u_iu_j$$

$$\frac{\partial J}{\partial u_i}=\sum_{j=1}^{n}a_{ij}u_j-b_i=0,\ i=1,\ \cdots,\ n \qquad (13.4.13)$$

按照上述特殊的离散方法，这里所得的结果与 §13.3 中的(13.3.20)是一致的。这时 A 的对称性以及 A 的半正定性即 $J_0(u_1,\ \cdots,\ u_n)$ 恒 $\geqslant 0$ 仍然成立。不同之点仅在于：当 $J_0(u_1,\ \cdots,\ u_n)=0$ 不能得到结论 $u_1=\cdots=u_n=0$ 而只能得到较弱的结论 $u_1=\cdots=u_n=c=$ 常数，因此矩阵 A 只是退化半正定。正如在 §13.4 末段所述，协调条件 (13.3.18) 是边值问题 (13.3.15～16) 有解的充要条件，因此也是变分问题(13.4.12)有解的充要条件。在离散化后，退化代数方程组(13.4.13)有解的充要条件是(13.3.23)。因此在用变分方法离散化时也要注意到保证协调条件在离散的意义下成立，即 (13.3.23) 成立，以使相应的代数方程组有解。

一般说来，椭圆算子具有对称正定性(或半正定性)，这是椭圆问题的物理特征的一种数学反映。用变分原理(或守恒原理)离散化时得到对称正定的系数阵，保持了这一特征，因此

是可取的。这对于数值解算也是有利的，因为关于对称正定矩阵的数值解法比较成熟(见§13.5及第八章)。应该指出，单纯用数值微分来离散化时，往往不能保证系数矩阵的对称正定性。

值得指出，在变分方法中，一旦能量积分写出后，离散化即按统一的程式进行。对于第二、三类即自然边界条件不需再作任何特殊处理，它们自动地得到保证。对于变系数而系数有间断时的交界条件也是这样。在变系数 $\beta=\beta(x, y)$ 时，在变分方法中，只须在每个单元 $C_{i+\frac{1}{2},j+\frac{1}{2}}$ 上取 β 为适当的常数平均值 $\beta_{i+\frac{1}{2},j+\frac{1}{2}}$，而单元能量积分离散化为

$$J_{i+\frac{1}{2},j+\frac{1}{2}}=\iint\limits_{C_{i+\frac{1}{2},j+\frac{1}{2}}}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy$$

$$\approx\frac{1}{4}\beta_{i+\frac{1}{2},j+\frac{1}{2}}[(u_{i+1,j}-u_{i,j})^2+(u_{i+1,j+1}-u_{i,j+1})^2+(u_{i,j+1}-u_{i,j})^2+(u_{i+1,j+1}-u_{i+1,j})^2]$$

$$-\frac{h^2}{4}(f_{i,j}u_{i,j}+f_{i+1,j}u_{i+1,j}+f_{i,j+1}u_{i,j+1}+f_{i+1,j+1}u_{i+1,j+1})$$

这样，当介质系数有间断(通常是分片为常数)而其间断线与格网线一致，交界条件也自能得到保证而无须特作处理，这也是变分方法的优点。这一切都是由于使能量达到极值的函数自动满足相应的微分方程和全部自然边界条件。

当定解区域的形状和介质间断线的分布很不规则时，则上述矩形格网剖分就不尽适应，而要采取三角形格网等其它的剖分形式而导致有限元法，这是更系统更通用的变分离散化方法，见第十四章。

§13.5 松 弛 法

椭圆差分方程组的主要数值解法是迭代法。这是因为未知数的个数比较多，往往有上百个或上千个，而且差分格式具有较高的规律性和重复性，迭代法的程序实现比较简单，存储量和运算量都比较省。在迭代法中又以松弛法最简单最常用，因此下面主要介绍这一种。当然，对于椭圆差分方程，除了松弛法外还有其它有效的迭代解法和直接解法，可以参考第八章以及[1]，[2]，[3]。

13.5.1 简单迭代法和松弛法

我们将以椭圆差分方程模型问题(13.2.14～15)为例来说明松弛法。取单位正方形域上的波瓦松方程第一类边值问题

$$\Omega:\ -\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=f,\ 0<x,\ y<1 \tag{13.5.1}$$

$$\Gamma:\ u=g \tag{13.5.2}$$

作等距方形格网，命

$$u_{i,j}=u(ih,\ jh),\ i,\ j=0,\ 1,\ \cdots,\ N, h=\frac{1}{N} \tag{13.5.3}$$

在内点取标准的五点差分格式

$$4u_{i,j}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}=h^2f_{i,j},\ i,\ j=1,\ \cdots,\ N-1 \tag{13.5.4}$$

在边点取已知值

$$u_{i,j}=g_{i,j},\ 当\ i\ 或\ j=0\ 或\ N \tag{13.5.5}$$

任取初值即初始场 $u_{i,j}^{(0)}$, $i, j=0, \cdots, N$ 满足边界条件

$$u_{i,j}^{(0)}=g_{i,j}, \quad \text{当 } i \text{ 或 } j=0 \text{ 或 } N \tag{13.5.6}$$

在内点按下列方式

$$4u_{i,j}^{(m+1)}-u_{i-1,j}^{(m)}-u_{i+1,j}^{(m)}-u_{i,j-1}^{(m)}-u_{i,j+1}^{(m)}=h^2f_{i,j}$$

即

$$u_{i,j}^{(m+1)}=\frac{1}{4}(h^2f_{i,j}+u_{i-1,j}^{(m)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m)}+u_{i,j+1}^{(m)}), \quad i,j=1, \cdots, N-1 \tag{13.5.7}$$

进行迭代，边点值则保持不变

$$u_{i,j}^{(m+1)}=u_{i,j}^{(m)} \quad i \text{ 或 } j=0 \text{ 或 } N \tag{13.5.8}$$

这叫做简单迭代法。这个迭代过程是收敛的，即不论初值 $u_{i,j}^{(0)}$ 怎样取，当 $m\to\infty$ 时 $u_{i,j}^{(m)}$ 恒收敛于差分方程组(13.5.4～5)的真解 $u_{i,j}$(见 13.5.4 节)。缺点是收敛很慢，实际上很少采用，但是在它的基础上可以导出种种有效的迭代解法。

注意在简单迭代法中，在每点产生新值时不能立即冲掉旧值，因为后者在其后邻接点工作时还要用到。可以把上法稍加变形，在每点产生新值时立即用以冲掉旧值，也就是说在每个工作点上充分利用已经产生的新值，这就得到所谓松弛法，也叫塞德尔(Seidel)迭代法。由于更新值的提前使用，松弛法收敛可比简单迭代法加快约一倍 (13.5.4 节)，而且程序实现上更方便，存储要求最低，对未知数 $u_{i,j}$ 只需一片场(以及存放 $f_{i,j}$ 的另一片场)。

在迭代时规定一个扫描顺序，比方说取行列顺序，即先按行号 $i=1, 2, \cdots$，每行内按列号 $j=1, 2, \cdots$。这样，在节点(i, j)工作时其邻点$(i-1, j)$及$(i, j-1)$都已有了新值，因此松弛法可以表为

$$u_{i,j}^{(m+1)}=\frac{1}{4}(h^2f_{i,j}+u_{i-1,j}^{(m+1)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m+1)}+u_{i,j+1}^{(m)}) \quad i, j=1, \cdots, N-1$$

可以将上式改写成增量的形式

$$u_{i,j}^{(m+1)}=u_{i,j}^{(m)}+\frac{1}{4}(h^2f_{i,j}+u_{i-1,j}^{(m+1)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m+1)}+u_{i,j+1}^{(m)}-4u_{i,j}^{(m)})$$

$$i, j=1, \cdots, N-1 \tag{13.5.9}$$

这里每点所给的增量 $\frac{1}{4}(\cdots)$就是要求方程局部达到平衡时应补充的量。还可以引进一个迭代参数即松弛因子 ω 而将上法推广为

$$u_{i,j}^{(m+1)}=u_{i,j}^{(m)}+\frac{\omega}{4}(h^2f_{i,j}+u_{i-1,j}^{(m+1)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m+1)}+u_{i,j+1}^{(m)}-4u_{i,j}^{(m)})$$

$$i, j=1, \cdots, N-1 \tag{13.5.10}$$

取 $\omega=1$ 时就是原来的塞德尔法。

在简单迭代法中，每次迭代的结果显然不依赖于扫描的次序。松弛法则不然，不同的扫描顺序可以导致不同(也可能相同)的结果，因此得到不同的松弛方案。除了上面所说的行列顺序(如图 13.7，这是比较简单常用的)外，还可以采取，比如说，奇偶顺序。把节点(i, j)按下标和 $i+j=$奇数或偶数分为奇偶两组。每次迭代时先扫遍奇点，后扫遍偶点(如图 13.8)。注意奇偶两组点作插花状分布，在奇点工作时，它的邻接点都是偶点，都只有旧值。其后在偶点工作时则邻接点都是奇点，都有了新值，因此松弛格式可以表为

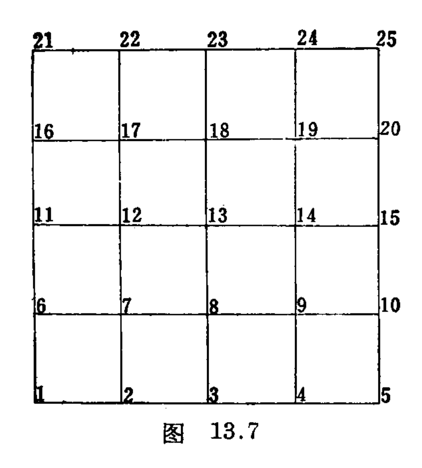


图 13.7

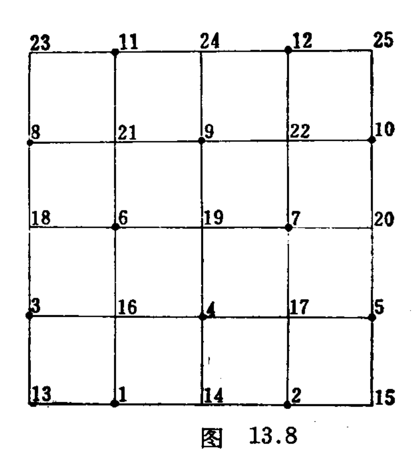


图 13.8

$$\left.\begin{aligned}
u_{i,j}^{(m+1)}&=u_{i,j}^{(m)}+\frac{\omega}{4}(h^2f_{i,j}+u_{i-1,j}^{(m)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m)}+u_{i,j+1}^{(m)}-4u_{i,j}^{(m)}),\ i+j=\text{奇}\\
u_{i,j}^{(m+1)}&=u_{i,j}^{(m)}+\frac{\omega}{4}(h^2f_{i,j}+u_{i-1,j}^{(m+1)}+u_{i+1,j}^{(m+1)}+u_{i,j-1}^{(m+1)}+u_{i,j+1}^{(m+1)}-4u_{i,j}^{(m)}),\ i+j=\text{偶}
\end{aligned}\right\}\quad(13.5.11)$$

注意在程序实现时，不论采用什么顺序，松弛法都表为相同的动态公式，命 $u(i, j)$ 表示场 $u_{i,j}$，$b(i, j)$ 表示场 $h^2f_{i,j}$，则有

$$\begin{aligned}
&u(i, j)+\frac{\omega}{4}(b(i, j)+u(i-1, j)+u(i+1, j)+u(i, j-1)\\
&\quad+u(i, j+1)-4u(i, j))\Rightarrow u(i, j)
\end{aligned}\qquad(13.5.12)$$

只是扫描控制有所不同。

在松弛法中，由于增加了选择参数 ω 的余地，以及（在较小程度上）选择扫描顺序的余地，方法的潜力增加了。可以证明（13.5.4 节），当松弛因子取在 $0<\omega<2$ 内时都是收敛的。取 $\omega>1$ 时叫做超松弛或过量松弛，即每点给予的增量超过使方程达到局部平衡之所需。反之，当 $\omega<1$ 时则叫做低松弛或欠量松弛。重要的是松弛因子 ω 有一个最优的选择 $\omega=\omega^*$ 能使收敛大大加快，这个最优值在 $1<\omega^*<2$ 之间，属于超松弛的范围。通常把这类方法叫做超松弛法，简称为松弛法。对于不同顺序，最优参数值 ω^* 以及最优的终极收敛速度是相同的，但初期的有限次内的收敛性能则有所不同。

将各点的未知量 $u_{ij}(i, j=1, \cdots, N-1)$ 按一定的扫描顺序（例如图 13.7 或图 13.8）排列，记为 $u_1, \cdots, u_n$，$n=(N-1)^2$。与此同时，将各内点的差分方程(13.5.4)也按同一顺序排列，得到线代数方程组

$$\sum_{j=1}^{n}a_{ij}u_j=b_i\quad i=1, \cdots, n\qquad(13.5.13)$$

于是简单迭代法表为

$$u_i^{(m+1)}=\frac{1}{a_{ii}}\left(b_i-\sum_{j\neq i}a_{ij}u_j^{(m)}\right)\qquad(13.5.14)$$

而松弛法则表为

$$u_i^{(m+1)}=u_i^{(m)}+\frac{\omega}{a_{ii}}\left(b_i-\sum_{j<i}a_{ij}u_j^{(m+1)}-\sum_{j\geq i}a_{ij}u_j^{(m)}\right)\qquad(13.5.15)$$

这里，对 u_i 加工时，按顺序在 i 以前的量 u_j 都取新值，故记为 $u_j^{(m+1)}$，在 i 及以后的分量则均取旧值，记为 $u_j^{(m)}$。

注意在不同的顺序下，系数 a_{ij}，b_i 是不同的，见 §14.5.3，相差的是行与列之间的排列。

命

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \quad b=\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \quad u=\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

方程组(13.5.13)表为矩阵形式

$$\boldsymbol{A}\boldsymbol{u}=\boldsymbol{b} \tag{13.5.16}$$

把 $\boldsymbol{A}$ 分解为对角线 $\boldsymbol{D}$ 和下三角 $\boldsymbol{L}$ 及上三角 $\boldsymbol{R}$ 三个部分

$$\boldsymbol{A}=\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{R} \tag{13.5.17}$$

$$\boldsymbol{D}=\begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix} \quad \boldsymbol{L}=\begin{bmatrix} 0 & & & \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \quad \boldsymbol{R}=\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{bmatrix}$$

于是简单迭代法可以表为

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{D}^{-1}(\boldsymbol{b}-(\boldsymbol{A}-\boldsymbol{D})\boldsymbol{u}^{(m)})$$

即

$$\boldsymbol{D}\boldsymbol{u}^{(m+1)}=-(\boldsymbol{A}-\boldsymbol{D})\boldsymbol{u}^{(m)}+\boldsymbol{b} \tag{13.5.18}$$

松弛法则表为

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m)}+\omega\boldsymbol{D}^{-1}(\boldsymbol{b}-\boldsymbol{L}\boldsymbol{u}^{(m+1)}-(\boldsymbol{D}+\boldsymbol{R})\boldsymbol{u}^{(m)})$$

即

$$(\boldsymbol{L}+\omega^{-1}\boldsymbol{D})\boldsymbol{u}^{(m+1)}=-(\boldsymbol{R}+(1-\omega^{-1})\boldsymbol{D})\boldsymbol{u}^{(m)}+\boldsymbol{b} \tag{13.5.19}$$

13.5.2 迭代法概述

上述几种迭代可以在形式上统一起来。设将系数阵 $\boldsymbol{A}$ 作适当的“分解”

$$\boldsymbol{A}=\boldsymbol{B}+\boldsymbol{C} \tag{13.5.20}$$

于是方程 $\boldsymbol{A}\boldsymbol{u}=\boldsymbol{b}$ 可表为

$$\boldsymbol{A}\boldsymbol{u}=\boldsymbol{B}\boldsymbol{u}+\boldsymbol{C}\boldsymbol{u}=\boldsymbol{b} \tag{13.5.21}$$

据此制定迭代法

$$\boldsymbol{B}\boldsymbol{u}^{(m+1)}+\boldsymbol{C}\boldsymbol{u}^{(m)}=\boldsymbol{b} \tag{13.5.22}$$

命

$$\boldsymbol{E}=-\boldsymbol{B}^{-1}\boldsymbol{C} \tag{13.5.23}$$

于是(13.5.21～22)分别表为

$$\boldsymbol{u}=\boldsymbol{E}\boldsymbol{u}+\boldsymbol{B}^{-1}\boldsymbol{b} \tag{13.5.24}$$

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{E}\boldsymbol{u}^{(m)}+\boldsymbol{B}^{-1}\boldsymbol{b} \tag{13.5.25}$$

矩阵 $\boldsymbol{E}=-\boldsymbol{B}^{-1}\boldsymbol{C}$ 在方法中起着关键的作用，叫做这个迭代法的迭代矩阵或收束矩阵。

回顾(13.5.18～19)可知简单迭代法相当于取

$$\boldsymbol{B}=\boldsymbol{D},\ \boldsymbol{C}=\boldsymbol{A}-\boldsymbol{D},\ \boldsymbol{E}=-\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{D})=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A}\xlongequal{\text{记作}}\boldsymbol{G} \tag{13.5.26}$$

而在松弛法相当于取

$$\boldsymbol{B}=\boldsymbol{L}+\omega^{-1}\boldsymbol{D},\ \boldsymbol{C}=\boldsymbol{R}+(1-\omega^{-1})\boldsymbol{D}$$

$$E=-(L+\omega^{-1}D)^{-1}(R+(1-\omega^{-1})D)\xlongequal{\text{记作}}H_\omega \tag{13.5.27}$$

当 $\omega=1$ 即塞德尔法则为

$$B=L+D,\ C=R,\ E=-(L+D)^{-1}R\xlongequal{\text{记作}}H_1 \tag{13.5.28}$$

对于迭代法首先要求收敛。这就是说，不论取什么初值 $u^{(0)}$，当 $m\to\infty$ 时 $u^{(m)}$ 应收敛于真解 u。定义误差向量

$$e^{(m)}=u^{(m)}-u,\quad m=0,\ 1,\ \cdots \tag{13.5.29}$$

收敛性就意味着：不论初始误差 $e^{(0)}$如何，当 $m\to\infty$ 时恒有 $e^{(m)}\to 0$。将 (13.5.25) 与 (13.5.24) 相减，得到

$$e^{(m+1)}=Ee^{(m)},\quad m=0,\ 1,\ \cdots \tag{13.5.30}$$

由此递推 $e^{(1)}=Ee^{(0)}$, $e^{(2)}=Ee^{(1)}=E^2e^{(0)}$, $\cdots$, 因此

$$e^{(m)}=E^me^{(0)},\quad m=0,\ 1,\ \cdots \tag{13.5.31}$$

矩阵 $E=-B^{-1}C$ 在每步迭代时起着压缩或放大误差的作用，因此叫做收束阵。方法是否收敛，快慢如何，主要取决于矩阵 E 的本征值模量的大小。

命 E 的全部本征值为 $\eta_1,\ \cdots,\ \eta_n$，相应地有本征向量 $w^{(1)},\ \cdots,\ w^{(n)}$，即 $Ew^{(i)}=\eta_iw^{(i)}$。如果这些本征向量是线性无关的，则任意 $e^{(0)}$ 必定可表为

$$e^{(0)}=\alpha_1w^{(1)}+\cdots+\alpha_nw^{(n)}$$

于是

$$e^{(1)}=Ee^{(0)}=\alpha_1\eta_1w^{(1)}+\cdots+\alpha_n\eta_nw^{(n)}$$

$$e^{(m)}=E^me^{(0)}=\alpha_1\eta_1^mw^{(1)}+\cdots+\alpha_n\eta_n^mw^{(n)}$$

显然可见，$e^{(m)}\to 0$ 的充要条件是 E 的本征值的绝对值都小于 1

$$|\eta_1|<1,\cdots,\ |\eta_n|<1 \tag{13.5.32}$$

事实上，可以证明，这一命题对于一般的矩阵 E（即使它不具有线性无关的本征向量组）也是成立的。通常把任意矩阵 E 的本征值的绝对值的最大值叫做 E 的谱半径，记为

$$\rho(E)=\max_{1\leqslant i\leqslant n}|\eta_i| \tag{13.5.33}$$

于是迭代法 (13.5.25) 收敛的充要条件就是收束阵 E 的谱半径 $\rho(E)<1$。

在收敛的情况，收敛快慢取决于 E 的本征值的绝对值的大小，基本上取决于本征值的最大绝对值即谱半径 $\rho=\rho(E)$ 的大小，愈小收敛愈快。通常取

$$R=-\ln\rho,\quad 即\quad \rho=e^{-R} \tag{13.5.34}$$

作为收敛快慢的度量，R 叫做渐近收敛速度。当迭代到一定次数 m，基本上稳化时，可以证明，任取一种模量（见第八章或 § 13.5.6）必有

$$\|e^{(m+1)}\|\approx\rho\|e^{(m)}\|,\quad \|e^{(m+k)}\|\approx\rho^k\|e^{(m)}\|$$

因此要把误差在原来的基础 $\|e^{(m)}\|$ 上压缩 ε 倍——ε 是小量，$\varepsilon<1$——所需的迭代次数 k 近似地满足

$$\rho^k=e^{-Rk}\approx\varepsilon$$

即

$$k\approx\frac{1}{R}\ln\left(\frac{1}{\varepsilon}\right) \tag{13.5.35}$$

例如，取 $\varepsilon=10^{-s}$，即要把误差在原有基础上压缩 10^{-s} 倍，相当于要提高十进制 s 位精度则需迭代次数为

$$k\approx\frac{1}{R}\ln 10^{s}\approx\frac{s}{R}\ln 10\approx 2.3s/R$$

实际迭代时需要给出某种收敛的判据，例如使误差向量 $e^{(m)}$ 的某种度量小于某个预给的容差以便结束迭代。但是误差 $e^{(m)}=u^{(m)}-u$ 本身是不知道的，不过相邻两次迭代的差额即“增量”

$$d^{(m+1)}=u^{(m+1)}-u^{(m)} \tag{13.5.36}$$

或者“余量”

$$r^{(m)}=b-Au^{(m)} \tag{13.5.37}$$

则是知道的。它们与误差 $e^{(m)}$ 的关系是

$$d^{(m+1)}=e^{(m+1)}-e^{(m)} \tag{13.5.38}$$

$$r^{(m)}=-Ae^{(m)},\quad e^{(m)}=-A^{-1}r^{(m)} \tag{13.5.39}$$

并且有

$$d^{(m+1)}=Ed^{(m)} \tag{13.5.40}$$

$$r^{(m+1)}=AEA^{-1}r^{(m)} \tag{13.5.41}$$

实践上可以采取

$$\|r^{(m)}\|=\max_{1<i<n}|r_i^{(m)}|\leqslant\rho \tag{13.5.42}$$

或者

$$\|d_i^{(m)}\|=\max_{1<i<n}|d_i^{(m)}|\leqslant\delta \tag{13.5.43}$$

作为收敛控制，ρ 或 δ 为预给的容差，对此将在 § 13.6 中讨论。

13.5.3 模型问题的频谱和矩阵表达

为了对迭代法进行分析，最好知道模型问题(13.5.5～6)的“频谱”即本征值分布。对应于差分边值问题(13.5.5～6)有差分本征值问题

$$\begin{cases}4u_{i,j}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}=\lambda u_{i,j},\quad i,\ j=1,\ \cdots,\ N-1 & (13.5.44)\\ u_{i,j}=0,\ \text{当}\ i\ \text{或}\ j=0\ \text{或}\ N & (13.5.45)\end{cases}$$

对于 $p,\ q=1,\ \cdots,\ N-1$ 命

$$\lambda^{(p,q)}=2\left(1-\cos\frac{p\pi}{N}\right)+2\left(1-\cos\frac{q\pi}{N}\right)=4\sin^2\frac{p\pi}{2N}+4\sin^2\frac{q\pi}{2N} \tag{13.5.46}$$

$$u_{i,j}^{(p,q)}=\sin\frac{pi\pi}{N}\sin\frac{qj\pi}{N},\quad i,\ j=0,\ \cdots,\ N \tag{13.5.47}$$

它们满足方程组(13.5.44～45)，因此 $\lambda^{(p,q)}$ 是本征值，都是正的，$u_{i,j}^{(p,q)}$ 是相应的本征“函数”，共有 $(N-1)^2$ 个。特别有最小及最大本征值

$$\lambda_{\min}=\lambda^{(1,1)}=4\left(1-\cos\frac{\pi}{N}\right)=8\sin^2\frac{\pi}{2N}\approx\frac{2\pi^2}{N^2} \tag{13.5.48}$$

$$\lambda_{\max}=\lambda^{(N-1,N-1)}=4\left(1-\cos\frac{(N-1)\pi}{N}\right)=8\cos^2\frac{\pi}{2N}=8-\lambda_{\min}\approx 8 \tag{13.5.49}$$

不难验证

$$\frac{1}{2}[\lambda^{(p,q)}+\lambda^{(N-p,N-q)}]=4,\quad p,\ q=1,\ \cdots,\ N-1$$

因此，所有的本征值对应于 $\lambda=4$，即在 $\lambda=4$ 的左右成对出现。

将 $u_{i,j}$ $(i,\ j=1,\ \cdots,\ N-1)$ 按一定顺序，例如按行列顺序 $u_{11},\ u_{12},\ \cdots,\ u_{21},\ u_{22},\ \cdots,$

$u_{N-1,N-1}$ 记为 $u_1, u_2, \cdots, u_n$, $n=(N-1)^2$, 得到列向量 $\boldsymbol{u}$, 类似地将 $f_{i,j}$ (连同边界值 $g_{i,j}$)排列为 $b_1, b_2, \cdots, b_n$ 得列向量 $\boldsymbol{b}$, 将方程 (13.5.4) 以及(13.5.42)也按相同顺序排列。于是边值问题(13.5.4～5)和本征值问题(13.5.44～45)分别表为

$$\sum_{q=1}^{n} a_{pq}u_q=b_p,\ \sum_{q=1}^{n} a_{pq}u_q=\lambda u_p,\ \ p=1, \cdots, N-1 \tag{13.5.50}$$

即

$$\boldsymbol{Au}=\boldsymbol{b},\ \ \boldsymbol{Au}=\lambda \boldsymbol{u} \tag{13.5.51}$$

这里 $\boldsymbol{A}=[a_{pq}]$是三对角块带状阵,阶数为 $n=(N-1)^2$

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{H} & -\boldsymbol{I} & & & \\ -\boldsymbol{I} & \boldsymbol{H} & -\boldsymbol{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\boldsymbol{I} & \boldsymbol{H} & -\boldsymbol{I} \\ & & & -\boldsymbol{I} & \boldsymbol{H} \end{bmatrix},\quad \boldsymbol{H}=\begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix} \tag{13.5.52}$$

对角块 $\boldsymbol{H}$ 是 $N-1$ 阶的三对角带状阵,非对角块上 $\boldsymbol{I}$ 表示 $N-1$ 阶单位阵。$\lambda^{(p,q)}$, $\boldsymbol{u}^{(p,q)}$ 也就是矩阵 $\boldsymbol{A}$ 的本征值和本征向量

$$\boldsymbol{A}\boldsymbol{u}^{(p,q)}=\lambda^{(p,q)}\boldsymbol{u}^{(p,q)},\ \ p, q=1, \cdots, N-1$$

因此阵 $\boldsymbol{A}$ 的性态数,即条件数,即最大与最小本征值的比为:

$$\begin{aligned} p=p(\boldsymbol{A}) &=\lambda_{\max}/\lambda_{\min}=(8\cos^2\pi/2N)/(8\sin^2\pi/2N) \\ &=\operatorname{ctg}^2(\pi/2N)\approx 4N^2/\pi^2 \end{aligned} \tag{13.5.53}$$

上列系数阵 $\boldsymbol{A}$ 的块状结构形式是由行列顺序以及五点差分格式的特点决定的。事实上, 取行列顺序时, 相当于把内节点 (i, j) 按行分为 $N-1$ 组 $S_1, S_2, \cdots, S_{N-1}$, 并对节点 (i, j)赋以统一的单标号 $p=1, 2, \cdots, n=(N-1)^2$。

$$S_1=\{(1, 1), (1, 2), \cdots, (1, N-1)\}=\{1, 2, \cdots, N-1\}$$
$$S_2=\{(2, 1), (2, 2), \cdots, (2, N-1)\}=\{N, N+1, \cdots, 2N-2\}$$

....

设节点的单标号 p 属于 S_k, 即该点位于第 k 行, 该点的差分方程至多与三个组 S_{k-1}(前一行), S_k(本行), S_{k+1}(后一行) 的节点相干。因此对于单标号 q 不属于 S_{k-1}, S_k, S_{k+1} 的 $a_{pq}=0$。因此 $\boldsymbol{A}$ 必为三对角块带状

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & & & \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{A}_{k,k-1} & \boldsymbol{A}_{kk} & \boldsymbol{A}_{k,k+1} \\ & & & \ddots & \ddots & \ddots \end{bmatrix} \tag{13.5.54}$$

块 $\boldsymbol{A}_{kl}$ 表示组 S_k(即第 k 行)与 S_l(即第 l 行)之间的相干性。此外, 在每行中节点 p 的差分方程至多与三点即 $p-1$(前一点), p (本点), $p+1$(后一点)相干,因此对角块是三对角线带状阵如(13.5.52)中的 $\boldsymbol{H}$; 再由相邻行间的相干性可知非对角块 $\boldsymbol{A}_{k,k-1}$, $\boldsymbol{A}_{k,k+1}$ 都是对角阵如(13.5.52)中的 $-\boldsymbol{I}$。

类似的方法可以定出在别的顺序下系数阵的表达式。在奇偶顺序下, 相当于把节点 (i, j)按照 $i+j=$奇数或偶数分为奇偶两组

$$S_1=\{(1, 2), (1, 4), \cdots, (2, 1), (2, 3), \cdots\}=\{1, 2, \cdots, m\}$$
$$S_2=\{(1, 1), (1, 3), \cdots, (2, 2), (2, 4), \cdots\}=\{m+1, m+2, \cdots, n\}$$

在五点差分式中，每点只与上下左右四个邻点相干，与斜向邻点不相干。因此在奇点的格式中只含一个奇点即本点和四个邻点都是偶点。偶点的格式则相反。因此，

$$A=\begin{bmatrix}\boldsymbol{D}_1 & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{D}_2\end{bmatrix} \tag{13.5.55}$$

这里对角块 $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ 分别表示奇组或偶组内部的相干性，本身都是对角阵。$\boldsymbol{A}_{12}$, $\boldsymbol{A}_{21}$ 表示奇对偶或偶对奇的相干性。在模型问题中，$\boldsymbol{D}_1$, $\boldsymbol{D}_2$ 的对角元都是 4, $\boldsymbol{A}_{12}$ 和 $\boldsymbol{A}_{21}$ 的每行至多有 4 个非零元素，都是 -1，细节从略。

不同顺序下的系数阵只相差行与列的重新排列，本征值是不变的，都是(13.5.46)，本征向量只相差分量标号的排列，其实还是同一个格网本征函数(13.5.47)。

在模型差分组系数阵(13.5.52)或(13.5.55)都是对称正定的。在复杂的区域以及变系数方程的情况，用变分原理或守恒原理推导的差分方程组能保证系数阵的对称正定性，见§13.2, §13.3。

13.5.4 收敛性分析

在模型问题中 $\boldsymbol{A}$ 的对角部 $\boldsymbol{D}=\frac{1}{4}\boldsymbol{I}$, 因简单迭代法的收束阵

$$\boldsymbol{G}=-\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{D})=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A}=\boldsymbol{I}-\frac{1}{4}\boldsymbol{A} \tag{13.5.56}$$

因此 $\boldsymbol{G}$ 的本征值 μ 与 $\boldsymbol{A}$ 的本征值 λ 之间有线性关系

$$\mu=1-\frac{1}{4}\lambda \tag{13.5.57}$$

于是，根据(13.5.49)

$$\mu_{\max}=1-\frac{1}{4}\lambda_{\min}\geqslant 0 \tag{13.5.58}$$

$$\mu_{\min}=1-\frac{1}{4}\lambda_{\max}=\frac{1}{4}\lambda_{\min}-1=-\mu_{\max}\leqslant 0 \tag{13.5.59}$$

注意，由于 $\boldsymbol{A}$ 的本征值对称于 $\lambda=4$，故 $\boldsymbol{G}$ 的本征值对称于 $\mu=0$，即正负成对出现。因此 $\boldsymbol{G}$ 的本征值的最大模，即谱半径，今后记作 μ 就是

$$\mu=\rho(\boldsymbol{G})=\mu_{\max}=1-\frac{1}{4}\lambda_{\min}=1-2\sin^2\frac{\pi}{2N}=\cos\frac{\pi}{N}<1 \tag{13.5.60}$$

(当然设 $N>1$)。因此简单迭代恒收敛。当 $N\gg 1$ 时

$$\mu=\cos\frac{\pi}{N}\approx 1-\pi^2/2N^2 \tag{13.5.61}$$

因此有收敛速度

$$R=-\ln\mu\approx\pi^2/N^2 \tag{13.5.62}$$

松弛法的收敛分析在理论上要复杂些，这是因为它的收束阵

$$\boldsymbol{H}_\omega=-(\boldsymbol{L}+\omega^{-1}\boldsymbol{D})^{-1}(\boldsymbol{R}+(1-\omega^{-1})\boldsymbol{D}) \tag{13.5.63}$$

不对称，有复数本征值。下面，我们仅对模型问题引述一些主要结论(参考[2], [1])。

(1) 当迭代参数 ω 取在 $0<\omega<2$ 之内时，松弛法都是收敛的。这一结论对于对称正定阵 $\boldsymbol{A}$ 普遍成立。

(2) 松弛法收束阵 $\boldsymbol{H}_\omega$ 的本征值 η 与简单迭代的收束阵 $\boldsymbol{G}=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A}$ 的本征值 μ 之间有对应关系

$$(\eta+\omega-1)^2/\omega^2\eta=\mu^2 \tag{13.5.64}$$

$\boldsymbol{G}$ 的本征值 μ 都是实数，正负成对出现，都在 $(-1,\ 1)$ 之内；$\boldsymbol{H}_\omega$ 的本征值 η 可以有复数。

(3) 取 $\omega=1$ 即塞德尔法时上列关系简化为

$$\eta=\mu^2 \tag{13.5.65}$$

因此塞德尔法收束阵 $\boldsymbol{H}_1$ 的谱半径 η_1 为简单迭代法收束阵 $\boldsymbol{G}$ 的谱半径 μ 的平方

$$\eta_1=\rho(\boldsymbol{H}_1)=\mu^2,\quad \mu=\rho(\boldsymbol{G}) \tag{13.5.66}$$

因此塞德尔法收敛速度比简单迭代法快一倍

$$R_1=-\ln\eta_1=-\ln\mu^2=2(-\ln\mu)=2R$$

(4) 当 ω 按照下列公式取优选值 ω^* 时 $\boldsymbol{H}_\omega$ 的谱半径 η_ω 达到极小：

$$\omega^*=\frac{2}{1+\sqrt{1-\mu^2}},\quad \mu=\rho(\boldsymbol{G}) \tag{13.5.67}$$

$$\eta_{\omega^*}=\rho(\boldsymbol{H}_{\omega^*})=\omega^*-1=\frac{1-\sqrt{1-\mu^2}}{1+\sqrt{1-\mu^2}} \tag{13.5.68}$$

这个优选值在 $1<\omega^*<2$ 之间，属于超松弛范围。

(5) 当 $\omega<\omega^*$，$\boldsymbol{H}_\omega$ 按模最大的本征值是正实数即 $\eta_\omega=\eta_{\max}$，并且单本征值。当 $\omega\leqslant\omega^*$ 时 $\boldsymbol{H}_\omega$ 的全部本征值（其中有复的）按模均相等，模值为 $\eta_\omega=\omega-1$。$\eta_\omega=\rho(\boldsymbol{H}_\omega)$ 对参数 ω 的依赖曲线如图 13.9 在极小点 $\omega=\omega^*$ 的右侧 $(\omega>\omega^*)$ 以斜率 1 缓升，在左侧 $(\omega<\omega^*)$ 以斜率 ∞ 急升。

图 13.9

(6) 在不同的扫描顺序下，尽管系数阵 $\boldsymbol{A}$ 以及收束阵 $\boldsymbol{G}$, $\boldsymbol{H}_\omega$ 的形式不同，例如在行列顺序 $\boldsymbol{A}$ 形如(13.5.52)，在奇偶顺序 $\boldsymbol{A}$ 形如(13.5.55)，但是上述各点结论都同样成立。由于不同顺序下 $\boldsymbol{G}$ 的本征值即 μ 相同，因此参数的优选值 ω^* 以及渐近收敛速度均相同，只是初期的即有限次的迭代效果有所不同。

用矩阵 $\boldsymbol{D}^{-1}\boldsymbol{A}$ 的性态数 $p=p(\boldsymbol{D}^{-1}\boldsymbol{A})$ 来刻画各种迭代法的收敛速度比较方便。命 $\boldsymbol{D}^{-1}\boldsymbol{A}$ 的最小最大本征值为 $\xi_{\min}$, $\xi_{\max}$。$\boldsymbol{G}=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A}$ 的最小最大本征值为 $\mu_{\min}=-\mu$, $\mu_{\max}=\mu$。由于 $\boldsymbol{D}^{-1}\boldsymbol{A}=\boldsymbol{I}-\boldsymbol{G}$，故

$$\xi_{\min}=1-\mu_{\max}=1-\mu,\quad \xi_{\max}=1-\mu_{\min}=1+\mu \tag{13.5.69}$$

因此

$$p=p(\boldsymbol{D}^{-1}\boldsymbol{A})=\frac{\xi_{\max}}{\xi_{\min}}=\frac{1+\mu}{1-\mu} \tag{13.5.70}$$

对于模型问题 $\boldsymbol{D}^{-1}\boldsymbol{A}=\frac{1}{4}\boldsymbol{A}$，因 $\boldsymbol{D}^{-1}\boldsymbol{A}$ 与 $\boldsymbol{A}$ 有相同的性态数即

$$p=p(\boldsymbol{D}^{-1}\boldsymbol{A})=p(\boldsymbol{A})=\operatorname{ctg}^2\frac{\pi}{2N}\approx 4N^2/\pi^2 \tag{13.5.71}$$

实践上重要的是 $p\gg1$ 即 $N\gg1$ 的情况。不难算出

$$\mu=\frac{p-1}{p+1}\approx1-\frac{2}{p}\approx1-\pi^2/2N^2 \tag{13.5.72}$$

$$\eta_1=\left(\frac{p-1}{p+1}\right)^2\approx1-\frac{4}{p}\approx1-\pi^2/N^2 \tag{13.5.73}$$

$$\eta_{\omega*}=\frac{1-\sqrt{1-\mu^2}}{1+\sqrt{1-\mu^2}}=\left(\frac{\sqrt{p}-1}{\sqrt{p}+1}\right)^2\approx 1-\frac{4}{\sqrt{p}}\approx 1-2\pi/N \tag{13.5.74}$$

因此得到渐近收敛速度

简单迭代:

$$R=-\ln\mu\approx\frac{2}{p}\approx\pi^2/2N^2 \tag{13.5.75}$$

塞德尔法:

$$R_1=-\ln\eta_1\approx\frac{4}{p}\approx\pi^2/N^2 \tag{13.5.76}$$

优选超松弛:

$$R_{\omega*}=-\ln\eta_{\omega*}\approx\frac{4}{\sqrt{p}}\approx 2\pi/N \tag{13.5.77}$$

塞德尔法比简单迭代快一倍，已见前述，而优选超松弛比塞德尔法又快约 $2\pi^{-1}N\approx 0.6N$ 倍，这是很可观的改进。对于三维模型问题，由三维系数阵的性态数与二维的相同，收敛速度也相同。总的说来，上述几种迭代法的收敛速度都是受系数阵（更确切些说阵 $\boldsymbol{D}^{-1}\boldsymbol{A}$）的性态数 p 制约的，矩阵病态愈甚，收敛愈慢。简单迭代和塞德尔法的收敛速度都反比于 p，而优选超松弛法则反比于 $\sqrt{p}$。

以上结论虽然是对模型问题在行列或奇偶顺序下说的，事实上，在更广的范围内，即对于所谓具有“性质 A”的系数阵在所谓“一致顺序”下结论(1)～(6)都成立。有关于细节比较繁琐，从略，可以参看[2]。笼统地说上述结论对于二阶椭圆方程的五点差分格式在相当广泛的顺序下均适合，而实际上应用范围甚至还要宽些。但是，对于四阶的椭圆方程，如板、壳问题中的重调和方程的差分格式，则上述结论未必成立；即使成立，由于性态数 $p\approx O(N^4)$，收敛也是很慢的。目前看来，对这一类问题采取其它方法，例如直接法或直接法与迭代法相结合的共轭斜量法（见第八章）等为宜。

综上所述，由于松弛法的算法简单，存储要求低，并可优选参数使收敛大大加快，它是椭圆差分方程的主要数值解法之一，特别适用于二阶（二维或三维）椭圆型问题。它还可以推广为行松弛或块松弛即成组松弛法，也有类似的优选参数方法，可以参考[1, 2]。

13.5.5 变参数松弛法

超松弛法取优选参数 $\omega=\omega^*$ 时能得最快的（相对于其它的 ω 值）终极即渐近收敛速度。但是，实际表明，有时在迭代初期误差反而增长，只当迭代了一定次数后才下降。反之，取 $\omega=1$ 即塞德尔法时，虽然最终收敛速度是很慢的，但迭代初期误差下降较快，特别对随机性的初始误差是这样。因此希望能把两法结合起来，保持最优的渐近速度，同时加速初期的收敛性以期及早结束迭代。对此可以采用变参数的松弛法。

一个最简单的结合方案是迭代第一步取 $\omega=1$，而从第二步起恒取 $\omega=\omega^*$。结果相对单纯的优选松弛有所改善。可以证明，当取奇偶顺序时，误差的欧氏模量是单调下降的。

另一种较为有效的加速法是在奇偶顺序下进行的。设想把每步迭代先扫奇点后扫偶点的步骤看作两次迭代，在这个奇偶交替过程中每次采用不同的迭代参数 $\omega=\omega_1,\ \omega_2,\ \omega_3,\ \omega_4,\ \cdots$。这样(13.5.11)就改为:

$$\left.\begin{aligned}&u_{i,j}^{(2m+1)}=u_{i,j}^{(2m-1)}+\frac{\omega_{2m+1}}{4}(b_{i,j}+u_{i-1,j}^{(2m)}+u_{i+1,j}^{(2m)}+u_{i,j-1}^{(2m)}+u_{i,j+1}^{(2m)}-4u_{i,j}^{(2m-1)}),\\&\quad i+j=\text{奇}\\&u_{i,j}^{(2m+2)}=u_{i,j}^{(2m)}+\frac{\omega_{2m+2}}{2}(b_{i,j}+u_{i-1,j}^{(2m+1)}+u_{i+1,j}^{(2m+1)}+u_{i,j-1}^{(2m+1)}+u_{i,j+1}^{(2m+1)}-4u_{i,j}^{(2m)}),\\&\quad i+j=\text{偶}\end{aligned}\right\}\tag{13.5.78}$$

参数 ω_k 用下列递推公式产生

$$\omega_1=1,\ \omega_2=\frac{2}{2-\mu^2},\quad \omega_{k+1}=\frac{4}{4-\mu^2\omega_k},\quad k=1,\ 2,\ \cdots\tag{13.5.79}$$

μ 的意义同前 $\mu=\rho(\boldsymbol{G})$。这就是所谓切比晓夫加速法，也是切比晓夫迭代法的一种变形，其推导见 §13.7。

这里第一步取 $\omega=1$ 以发挥塞德尔迭代初期压缩误差的作用。当 $k\to\infty$ 时 ω_k 稳定到一个极限值。只须在递推公式中命 $\omega_{k+1}=\omega_k$ 就可以算这个极值就是超松弛的优选值

$$\omega_k\to\frac{2}{1+\sqrt{1-\mu^2}}=\omega^*\tag{13.5.80}$$

因此渐近收敛速度与优选松弛法同。至于初期收敛性则有显著改进，见 13.5.6 节这个方法在超松弛的基础稍作变化就可实现，因此在实践上是可取的。

13.5.6 初期收敛性的比较

本节对常参数和变参数松弛法的初期收敛性进行比较。为此目的，先提一下向量和矩阵模量的概念(参看第八章)。

通常衡量一个向量 $\boldsymbol{v}$ 的“大”“小”可以采取一定的模量(即范数) $\|\boldsymbol{v}\|$，例如：向量的欧几里得模量定义为

$$\|\boldsymbol{v}\|=(v_1^2+\cdots+v_n^2)^{1/2}\tag{13.5.81}$$

对于所取定的向量模量可以规定矩阵 $\boldsymbol{E}$ 的模量

$$\|\boldsymbol{E}\|=\max_{v\neq0}\frac{\|\boldsymbol{E}\boldsymbol{v}\|}{\|\boldsymbol{v}\|}\tag{13.5.82}$$

并且恒有不等式

$$\|\boldsymbol{E}\boldsymbol{v}\|\leqslant\|\boldsymbol{E}\|\cdot\|\boldsymbol{v}\|\tag{13.5.83}$$

这里等号是可以达到的，即总存在适当的向量 $\boldsymbol{v}$ 使得 $\|\boldsymbol{E}\boldsymbol{v}\|=\|\boldsymbol{E}\|\cdot\|\boldsymbol{v}\|$。可以证明，当向量取欧氏模量(13.5.81)时，相应的矩阵欧氏模量 $\|\boldsymbol{E}\|$ 就是矩阵 $\boldsymbol{E}^T\boldsymbol{E}$ 的谱半径的平方根即

$$\|\boldsymbol{E}\|=(\rho(\boldsymbol{E}^T\boldsymbol{E}))^{1/2}\tag{13.5.84}$$

当 $\boldsymbol{E}$ 为对称阵时，$\boldsymbol{E}^T=\boldsymbol{E}$，$\rho(\boldsymbol{E}^T\boldsymbol{E})=(\rho(\boldsymbol{E}))^2$，因此其欧氏模量与谱半径同即

$$\|\boldsymbol{E}\|=\rho(\boldsymbol{E})\tag{13.5.85}$$

在非对称阵则两者可以不等但恒满足不等式

$$\rho(\boldsymbol{E})\leqslant\|\boldsymbol{E}\|\leqslant n\rho(\boldsymbol{E})\tag{13.5.86}$$

迭代法的收敛条件是其收束阵 $\boldsymbol{E}$ 的谱半径 $\rho(\boldsymbol{E})<1$。这时虽然保证终极的收敛性，但在有限次迭代的误差增减可以比较复杂。事实上，根据(13.5.31)，(13.5.83)

$$\|\boldsymbol{e}^{(m)}\|\leqslant\|\boldsymbol{E}^m\|\cdot\|\boldsymbol{e}^{(0)}\|\tag{13.5.87}$$

当 $\boldsymbol{E}$ 为对称阵时，例如在简单迭代法就是这样，

$$\|\boldsymbol{E}\|=\rho(\boldsymbol{E}),\ \|\boldsymbol{E}^m\|=\rho(\boldsymbol{E}^m)=(\rho(\boldsymbol{E}))^m \tag{13.5.88}$$

当 $\rho(\boldsymbol{E})<1$ 时，误差按欧氏模量是单调递减的。但是，例如在超松弛法中，收束阵 $\boldsymbol{E}$ 不对称，虽然 $\rho(\boldsymbol{E})<1$ 而可能欧氏模量 $\|\boldsymbol{E}\|>1$，这时可以发生 $\|\boldsymbol{e}^{(1)}\|=\|\boldsymbol{E}\|\cdot\|\boldsymbol{e}^{(0)}\|>\|\boldsymbol{e}^{(0)}\|$，因此误差模量可能增长。但是

$$\|\boldsymbol{E}^m\|\leqslant n\rho(\boldsymbol{E}^m)=n(\rho(\boldsymbol{E}))^n\to 0 \quad (m\to\infty)$$

到了最终，模量还是递减至于零的。

对于优选超松弛法，收束阵为 $\boldsymbol{H}_{\omega*}$

$$\eta=\rho(\boldsymbol{H}_{\omega*})=\frac{2}{1+\sqrt{1-\mu^2}}<1 \tag{13.5.89}$$

$$\mu=\rho(\boldsymbol{G})=\frac{2\eta^{\frac{1}{2}}}{1+\eta}<1 \tag{13.5.90}$$

$$\boldsymbol{e}^{(m)}=\boldsymbol{H}_{\omega*}^m\boldsymbol{e}^{(0)},\quad \rho(\boldsymbol{H}_{\omega*}^m)=\eta^m \tag{13.5.91}$$

$$\|\boldsymbol{e}^{(m)}\|\leqslant\|\boldsymbol{H}_{\omega*}^m\|\cdot\|\boldsymbol{e}^{(0)}\| \tag{13.5.92}$$

虽然 $\eta=\rho(\boldsymbol{H}'_{\omega*})$ 和 $\eta^m=\rho(\boldsymbol{H}_{\omega*}^m)$ 都 <1，但当 m 较小时，$\|\boldsymbol{H}_{\omega*}\|$ 和 $\|\boldsymbol{H}_{\omega*}^m\|$ 都 >1。事实上，可以证明(见[2])

$$\varphi^*_{(m)}=\|\boldsymbol{H}_{\omega*}^m\|=\left[\frac{2m}{\mu}+\sqrt{\left(\frac{2m}{\mu}\right)^2+1}\,\right]\cdot\eta^m,\quad m=0,\ 1,\ \cdots \tag{13.5.93}$$

右端第一因子[…]基本上随 m 线性上升，第二因子 η^m 随 m 指数状下降。由于 $\eta\sim1$，$\mu\sim1$，当 m 较小时第一因子占主导，当 m 很大时第二因子占主导。因此，当 m 从 0 开始增大时，$\varphi(m)$ 从 $\varphi(0)=1$ 上升到一定的峰值(>1)然后单调下降以至于 0，这表示在迭代初期误差模量会上升。

对于 $\omega=1$ 即塞德尔法，可以证明[2]

$$\varphi_1(m)=\|\boldsymbol{H}_1^m\|=\sqrt{1+\mu^2}\,\mu^{2m-1} \tag{13.5.94}$$

μ 的意义同前。$\varphi_1(m)$ 是单调下降的。

对于切比雪夫加速法可以证明[2]有矩阵 $\boldsymbol{G}_1,\ \boldsymbol{G}_2,\ \cdots,\ \boldsymbol{G}_{2m}$（这里 $\boldsymbol{G}_{2m}$ 是与 $2m$ 次切氏多项式相连系的对称矩阵）使得

$$\boldsymbol{e}^{(m)}=\boldsymbol{G}_{2m}\boldsymbol{e}^{(0)}$$

并且

$$\varphi_c(m)=\|\boldsymbol{G}_{2m}\|=\frac{2}{1+\eta^{2m}}\eta^m \tag{13.5.95}$$

η 的意义同前，这也是单调下降的。

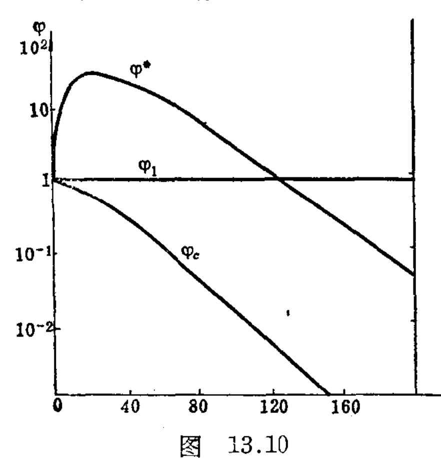


图 13.10

对于模型问题，取

$$N=128,\quad \mu=\cos\pi/128=0.99997$$
$$\omega^*=1.952,\quad \eta=0.952$$

曲线 $\varphi^*(m)$，$\varphi_1(m)$，$\varphi_c(m)$ 见图 13.10（参考[3]）。在优选超松弛法，最初 20 次迭代到 20 次时 $\varphi^*(m)$ 上升达 30 倍，至 128 次($\approx N$)开始复原。在塞德尔法，$\varphi_1(m)$ 微微向下的曲线，没有误差增大现象。最好的是切氏加速法 $\varphi_c(m)$，它位于其它两曲线之下，初期误差下降最快，大约 32 ($\approx N/4$)次后 $\varphi(m)\approx\rho^m$ 达到渐近收敛速度。

§13.6 实际计算中的处理

13.6.1 收敛控制和问题规模的估计

如13.5.2节所述,迭代过程的结束可以用余量控制

$$\|\boldsymbol{r}^{(m)}\|=\max_{1\leqslant i\leqslant n}|r_i^{(m)}|\leqslant\rho \tag{13.6.1}$$

或增量控制

$$\|\boldsymbol{d}^{(m)}\|=\max_{1\leqslant i\leqslant n}|d_i^{(m)}|\leqslant\delta \tag{13.6.2}$$

这里ρ或δ是预给的容差。容差过宽，精确度成问题，过严则浪费机器时间。关于究竟取余量还是取增量来控制以及容差如何定的问题，一般没有定论，要结合具体问题作具体分析。下面只能给出粗略的讨论。

首先,对于椭圆差分方程,迭代达到过高精度,既耗费大也不必要。这是因为,在用差商代替微商时必有截断误差。以模型问题为例,由于采用了二阶精度的数值微分公式,微分方程的精确解并不严格满足差分方程(13.5.4)而只是近似满足

$$f_{ij}-\frac{1}{h^2}(4u_{i,j}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1})\approx\frac{M_4}{b}h^2\approx O(h^2) \tag{13.6.3}$$

这可以从幂次展开得出,M_4为四阶导数的上界。也就是说

$$h^2f_{i,j}-(4u_{i,j}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1})\approx O(h^4) \tag{13.6.4}$$

注意向量$\boldsymbol{b}$相当于这里的$h^2f_{i,j}$而不是$f_{i,j}$。因此在迭代时只需将余量$\boldsymbol{r}^{(m)}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}^{(m)}$压至$O(h^4)\approx O(N^{-4})$的量级，再低就没有意义了。因此余量容差大致可取为$\rho\approx O(N^{-4})$。

如果用增量$d^{(m)}$控制,则容差δ应有所不同。这是由于

$$\begin{aligned}\boldsymbol{d}^{(m+1)}&=\boldsymbol{e}^{(m+1)}-\boldsymbol{e}^{(m)}=\boldsymbol{A}\boldsymbol{e}^{(m)}-\boldsymbol{e}^{(m)}=(\boldsymbol{A}-\boldsymbol{I})\boldsymbol{e}^{(m)}\\&=(\boldsymbol{A}-\boldsymbol{I})(-\boldsymbol{A}^{-1}\boldsymbol{r}^{(m)})=(\boldsymbol{A}^{-1}-\boldsymbol{I})\boldsymbol{r}^{(m)}\end{aligned} \tag{13.6.5}$$

在模型问题中$\boldsymbol{A}$的最小本征值为N^{-2}量级，即$\boldsymbol{A}^{-1}$的最大本征值为N^2量级，故用矩阵$(\boldsymbol{A}^{-1}-\boldsymbol{I})$作用时可能(最坏的情况)把向量放大$N^2$倍,因此要使$\boldsymbol{r}^{(m)}$压至$O(N^{-4})$量级应取增量$\boldsymbol{d}^{(m)}$的容差约为$\delta\approx O(N^{-6})$。

现以取余量控制$\rho\approx O(N^{-4})$为例,试对迭代次数以及计算总量作一粗略估计。由于$\boldsymbol{r}^{(m)}=\boldsymbol{A}\boldsymbol{e}^{(m)}$,

$$r_i^{(m)}=\sum_{j=1}^n a_{ij}e_j^{(m)},\ i=1,\ \cdots,\ n$$

$$|r_i^{(m)}|\leqslant\sum_{j=1}^n|a_{ij}|\cdot|e_j^{(m)}|\leqslant\max_j|e_j^{(m)}|\cdot\sum_{j=1}^n|a_{ij}|$$

$$|r_i^{(m)}|\leqslant\sum_{j=1}^n|a_{ij}|\cdot|e_j^{(m)}|\leqslant\max_j|e_j^{(m)}|\cdot\sum_{j=1}^n|a_{ij}|$$

因此

$$\|\boldsymbol{r}^{(m)}\|\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{e}^{(m)}\|,\ \|\boldsymbol{A}\|=\max_{1\leqslant i\leqslant n}\sum_{j=1}^n|a_{ij}| \tag{13.6.6}$$

在模型问题$\|\boldsymbol{A}\|$就是各点差分方程的系数绝对和的极大值,在模型问题$\|\boldsymbol{A}\|=8$,因此

$$\|\boldsymbol{e}^{(m)}\|\geqslant\frac{1}{8}\|\boldsymbol{r}^{(m)}\|$$

因此，当余量$\boldsymbol{r}^{(m)}$压至$O(N^{-4})$量级时，误差$\boldsymbol{e}^{(m)}$压至ε，$\varepsilon\geqslant cN^{-4}$，$c$为与$N$无关的常数。根据(13.5.35),迭代至此所需的次数

$$k \approx \frac{1}{R}\ln\left(\frac{1}{\varepsilon}\right) \approx \frac{1}{R}(4\ln N + \ln c^{-1}) \approx \frac{4}{R}\ln N \tag{13.6.7}$$

这里认为 N 充分大，略去了与 N 无关常数 $\ln c^{-1}$，R 为渐近收敛速度。

(13.6.7)只是迭代次数的上界估计，实践上，当有较好的初始值时，可以在远低于这个估计的次数就达到合理的结果。

设格网的节点总数为 M，迭代时每点的运算量为 φ，于是一次迭代的运算量为 $M\varphi$，迭代 k 次总的运算量 $\Phi = kM\varphi$。在二维模型问题，$M = N^2$，每点运算量 $\varphi \approx 6$，因此 $\Phi \approx 24N^2\ln N/R$。

优选松弛 $\omega = \omega^*$：$R \approx \dfrac{2\pi}{N}$，$\Phi \approx 2N^3\ln N$

塞德尔法 $\omega = 1$：$R \approx \dfrac{\pi^2}{N^2}$，$\Phi \approx 2.6N^4\ln N$

对于一般的二阶椭圆差分方程，这个界限估计在量级上也是对的。结合着机器的速度和容量，可以对问题的规模作出基本估计，以便在解题时心中大致有谱。

13.6.2 迭代参数的试选方法

通常在实践上 $N \gg 1$，对于不同参数 ω，收敛速度有数量级的差别，因此 ω 是否为优选影响很大。

收敛速度取决于收束阵 $\boldsymbol{H}_\omega$ 的谱半径 η_ω，后者对 ω 的依赖曲线如图 13.9。曲线在优选点 $\omega = \omega^*$ 有反折。在 $\omega < \omega^*$ 的方面，η_ω 以斜率 ∞ 急升；在 $\omega > \omega^*$ 的方面以斜率 1 缓升。因此在 ω^* 的邻近取参数 ω 宁可偏大而不要偏小。

优选值 ω^* 取决于谱半径 $\rho(\boldsymbol{G}) = \mu$。在模型问题中确知 $\mu = \cos\pi/N$；在一般情况下 μ 是不知道的，但可以在迭代过程中定出。在迭代的每一步，余量 $\boldsymbol{r}^{(m)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}^{(m)}$ 或增量 $\boldsymbol{d}^{(m+1)} = \boldsymbol{u}^{(m+1)} - \boldsymbol{u}^{(m)}$ 是现成的已知值。它们的发展规律是(14.5.40～41)即

$$\boldsymbol{r}^{(m+1)} = \boldsymbol{A}\boldsymbol{H}_\omega\boldsymbol{A}^{-1}\boldsymbol{r}^{(m)} \tag{13.6.8}$$

$$\boldsymbol{d}^{(m+1)} = \boldsymbol{H}_\omega\boldsymbol{d}^{(m)} \tag{13.6.9}$$

$\boldsymbol{A}\boldsymbol{H}_\omega\boldsymbol{A}^{-1}$ 与 $\boldsymbol{H}_\omega$ 有相同的本征值。为了估计 μ，开始迭代时取 $\omega = 1$，即 $\boldsymbol{H}_\omega = \boldsymbol{H}_1$。由于它的按模最大本征值 $\eta_1 = \rho(\boldsymbol{H}_1) = \mu^2$，并是为正实单根，因此，如以余量为基础，根据幂法(第十章)可以定出

$$\|\boldsymbol{r}^{(m+1)}\| / \|\boldsymbol{r}^{(m)}\| \to \eta_1 = \mu^2$$

也就是说，一直迭代到上列比值稳定为止。这里模量可以取为最大模 $\|\boldsymbol{r}\| = \max|r_i|$ 或欧氏模 $\|\boldsymbol{r}\| = (\Sigma r_i^2)^{1/2}$。得到 μ^2 后再按公式(13.5.67)计算 ω^*，以后就改用这个参数值或稍放大些继续进行迭代。也可以在开始时取 $\omega > 1$ 但 $< \omega^*$，这时 $\boldsymbol{H}_\omega$ 的按模最大本征值 η_ω 也是正实数单根，故按幂法也有

$$\|\boldsymbol{r}^{(m+1)}\| / \|\boldsymbol{r}^{(m)}\| \to \eta_\omega \tag{13.6.10}$$

再用公式(13.5.64)计算 μ^2，由此再用公式(13.5.67)计算 ω^*。注意开始迭代时的参数 ω 一定要小于 ω^*，否则 $\boldsymbol{H}_\omega$ 的按模最大本征值是复数，幂法不一定收敛。

也可以在增量 $\boldsymbol{d}^{(m)}$ 的基础上使用幂法定 μ^2，

$$\boldsymbol{d}^{(m+1)} = \boldsymbol{H}_\omega\boldsymbol{d}^{(m)}, \quad \|\boldsymbol{d}^{(m+1)}\| / \|\boldsymbol{d}^{(m)}\| \to \eta_\omega \tag{13.6.11}$$

但实践表明这样做的精度稍差。

在较为复杂的问题，前述优选参数的理论包括公式(13.5.67～68)不一定适用，这时可以用试算对比的方法来定出较优的因子(见第八章)。

13.6.3 关于复杂情况的处理

以上讲的二阶常系数方程矩形区域第一类边值问题。当边界条件由第一类变至第二、三类，区域由矩形变为不规则，微分方程由常系数变为变系数，二阶变为四阶时，情况就复杂化，处理上要采取相应措施，但基本原则是相仿的。

当边界条件为第二类，例如对于方程(13.5.1)区域的上部边界条件改为

$$\frac{\partial u}{\partial \nu}=q \tag{13.6.12}$$

于是边点(i, N)可以取，例如(13.3.5)即

$$2u_{i,N}-\frac{1}{2}u_{i-1,N}-\frac{1}{2}u_{i+1,N}-u_{i,N-1}=\frac{1}{2}h^2f_{i,N}+hq_{i,N}=b_{i,N} \tag{13.6.13}$$

而动态公式(13.5.12)要局部地改成

$$u(i,\ N)+\frac{\omega}{2}\Big[b(i,\ N)+\frac{1}{2}u(i-1,\ N)+\frac{1}{2}u(i+1,\ N)$$

$$+u(i.\ N-1)-2u(i,\ N)\Big]\Rightarrow u(i,\ N \tag{13.6.14}$$

因此边界点差分格式的多样化增加了程序的复杂性。

一个常见的特殊情况是单纯的第二类边值问题如(13.3.15～16)，当然给定f与g满足协调条件(13.3.18)以保证有解。按照§14.3或§13.4离散化后得到五点差分方程组(13.3.20)即

$$\sum_{j=1}^{n}a_{ij}u_j=b_i,\quad i=1,\ \cdots,\ n \tag{13.6.15}$$

矩阵$\boldsymbol{A}$是对称退化半正定，$\boldsymbol{A}$有本征值0，其它本征值为正，在离散化时应使协调条件(13.3.18)得到保持即应满足协调条件

$$\Sigma b_i=0 \tag{13.6.16}$$

以保证(13.6.15)有解。如果(13.6.16)不被严格满足，则如在§13.3末所述可对b_i稍加调整后使(13.6.16)成立。这样的差分方程组有解，但不唯一，不同解之间相差一个常数。一种习惯的解法是固定解的一个分量然后用松弛法。较好的办法则是直接用松弛法。对这一情况，松弛法的理论和实践基本上可以照搬，只有个别之点稍加修正。事实上，可以证明：松弛法的收束阵$\boldsymbol{H}_\omega$的本征值η与简单迭代的收束阵$\boldsymbol{G}$的本征值μ之间仍有关系(13.5.64)。不同之点只在于$\boldsymbol{B}$的最大模本征值为$\mu=\pm1$，对应的η有一个为1，另一个为$(\omega-1)^2$。当协调条件(13.6.16)成立时，对于$0<\omega<2$范围内的超松弛法对于任意初始向量都是收敛的，即收敛到差分方程组(13.6.15)的无穷多个解中的一个，而松弛因子ω的优选值则是

$$\omega^*=\frac{2}{1+\sqrt{1-\mu_2^2}} \tag{13.6.17}$$

形式与(13.5.67)相似，只是把那里的$\mu=\rho(G)$代以G的按模次大本征值μ_2。在迭代过程中试选ω^*的方法也同于13.6.2节，即任取$\omega<\omega^*$ (13.6.17)，进行迭代，无论余量或增量按幂法所得的模量比或分量比必趋向于H_ω的次大本征值η_2，由(13.5.64)可算出G的次大本征值μ_2，再由(13.6.17)算出ω^*。这个直接的最优超松弛法要比固定一个分量的最优

超松弛法收敛快得多。有关的情况可以参考[6]或第八章。

当区域具有不规则边界时,内存的安排和扫描控制会大大复杂化。在有些情况下可以把这种边界不规则的格网套在一个规则格网之内如图13.11,并把节点分为两类,一类是原有节点,是计算点,另一类是新添的虚点,是不计算点。迭代扫描时按外套规则格网进行,对于虚点是"空跑",这是付出一定的存储空间和计算时间的代价来换取程序的简化。

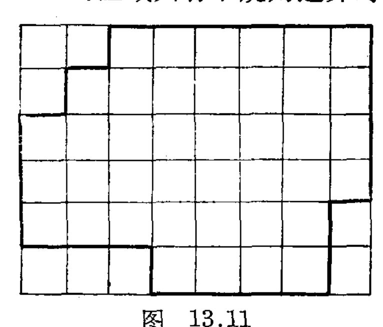


图 13.11

当微分方程为变系数时,差分格式也变复杂,例如成为

$$\sum_{p,q}\alpha_{i,j}^{(p,q)}u_{i+p,j+q}=b_{i,j} \tag{13.6.18}$$

在二阶方程的五点或九点格式,$p, q=0, \pm1$,在二阶微分方程以及四阶微分方程,则p, q的范围更宽。这时迭代的动态公式自然是

$$u(i, j)+\frac{\omega}{\alpha_{i,j}^{(0,0)}}\left[b(i, j)-\sum_{p,q}\alpha_{i,j}^{(p,q)}u(i+p, j+q)\right]\Rightarrow u(i, j) \tag{13.6.19}$$

这里需要把逐点的系数保存,使得存储的负担大大加重,例如在五点格式,要增加五片场来保存系数。也可以只列出$\alpha_{i,j}^{(p,q)}$的计算公式,而迭代时逐点临时计算值,利用计算机的速度以补偿存储量的不足。

一般说来,对于实际的计算问题,三个因素——一、存储量,二、计算工作量,三、程序复杂性(相当于准备工作量)必须权衡轻重,统一起来考虑。其中任一因素都可能构成客观上解题的限度。对于变系数以及几何上复杂的问题,第三个因素往往占主导地位,目前有限元法(第十四章)对此解决得比较好,使得多种多样的复杂问题在算法上基本统一起来,便于标准化,对于这类问题以采用有限元法为宜。

§13.7 变参数简单迭代法

13.7.1 简单迭代的加速

对于模型问题(13.5.4~5),仿照13.5.1节中从松弛法推广到超松弛法的途径,把简单迭代法(13.5.7)改写成

$$u_{i,j}^{(m+1)}=u_{i,j}^{(m)}+\frac{1}{4}(h^2f_{i,j}+u_{i-1,j}^{(m)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m)}+u_{i,j+1}^{(m)}-4u_{i,j}^{(m)}),\ i, j=1, \cdots, N-1$$

为了加速的目的,引进迭代参数ω,得到

$$u_{i,j}^{(m+1)}=u_{i,j}^{(m)}+\frac{\omega}{4}(h^2f_{i,j}+u_{i-1,j}^{(m)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m)}+u_{i,j+1}^{(m)}-4u_{i,j}^{(m)}) \tag{13.7.1}$$

简单迭代法相当于$\omega=1$。

把差分方程组写成代数的形式

$$\sum_{j=1}^{n}a_{ij}u_j=b_i,\quad i=1, \cdots, n \tag{13.7.2}$$

即

$$Au=b \tag{13.7.3}$$

上述方法就表为

$$u_i^{(m+1)}=u_i^{(m)}+\frac{\omega}{a_{ii}}\left(b_i-\sum_{j=1}^{n}a_{ij}u_j^{(m)}\right) \tag{13.7.4}$$

即

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m)}+\omega\boldsymbol{D}^{-1}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{u}^{(m)}) \tag{13.7.5}$$

亦即

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{G}_\omega\boldsymbol{u}^{(m)}+\omega\boldsymbol{D}^{-1}\boldsymbol{b} \tag{13.7.6}$$

$$\boldsymbol{G}_\omega=\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}$$

当 $\omega=1$ 时 $\boldsymbol{G}_\omega$ 就是简单迭代的收束阵 $\boldsymbol{G}$，记其谱半径为 μ，

$$\boldsymbol{G}_1=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A}=\boldsymbol{G},\quad \rho(\boldsymbol{G})=\mu \tag{13.7.7}$$

阵 $\boldsymbol{G}_\omega$ 的本征值 η 与阵 $\boldsymbol{D}^{-1}\boldsymbol{A}$ 的本征值 ξ 之间有线性关系

$$\eta=1-\omega\xi \tag{13.7.8}$$

在 $\xi\eta$ 平面上这是通过点 $(0,\ 1)$ 和 $(\omega^{-1},\ 0)$ 的直线。由于 ξ 在区间

$$0<1-\mu=\xi_{\min}\leqslant\xi\leqslant\xi_{\max}=1+\mu$$

内变，因此从图13.12(其中画出了相当于 $\omega^{-1}=\frac{1+\mu}{2}$ 及 $\omega^{-1}=1$ 的两条直线)可以看出：当取 $\frac{1+\mu}{2}<\omega^{-1}<\infty$ 时本征值 η 的最大模即谱半径 $\rho(\boldsymbol{G}_\omega)<1$，其中当取

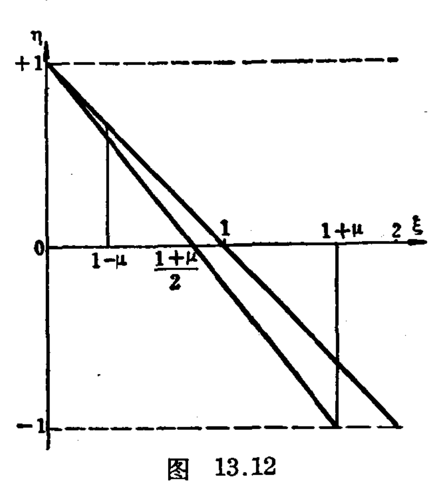


图 13.12

$$\omega^{-1}=\frac{1}{2}(\xi_{\min}+\xi_{\max})=\frac{1}{2}(1-\mu+1+\mu)=1$$

时 $\rho(\boldsymbol{G}_\omega)$ 极小，其值为 μ。这就是说取 $0<\omega<\frac{2}{1+\mu}$ 时迭代都收敛，其中收敛最快是取 $\omega=1$，就是普通的简单迭代。

这样，对简单迭代法引进参数 ω 后未能得到改进。但是，如果把参数改为可变的，

$$u_{i,j}^{(m+1)}=u_{i,j}^{(m)}+\frac{\omega_{m+1}}{4}(h^2f_{i,j}+u_{i-1,j}^{(m)}+u_{i+1,j}^{(m)}+u_{i,j-1}^{(m)}+u_{i,j+1}^{(m)}-4u_{i,j}^{(m)}),$$

$$i,\ j=1,\ \cdots,\ N-1 \tag{13.7.9}$$

并利用所谓切比雪夫多项式的性质后，则可有本质改进，将能得到与优选超松弛法相当的收敛速度，并有应用范围较宽的优点。

变参数简单迭代法(13.7.9)可以表为

$$\boldsymbol{u}^{(m+1)}=(\boldsymbol{I}-\omega_{m+1}\boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{u}^{(m)}+\omega_{m+1}\boldsymbol{D}^{-1}\boldsymbol{b} \tag{13.7.10}$$

为了方便，命

$$\hat{\boldsymbol{A}}=\boldsymbol{D}^{-1}\boldsymbol{A}=\boldsymbol{I}-\boldsymbol{G},\quad \hat{\boldsymbol{b}}=\boldsymbol{D}^{-1}\boldsymbol{b} \tag{13.7.11}$$

于是原来的方程组(13.7.3)等价于

$$\hat{\boldsymbol{A}}\boldsymbol{u}=\hat{\boldsymbol{b}} \tag{13.7.12}$$

也就是

$$\boldsymbol{u}=(\boldsymbol{I}-\omega_{m+1}\hat{\boldsymbol{A}})\boldsymbol{u}+\omega_{m+1}\hat{\boldsymbol{b}} \tag{13.7.13}$$

而迭代格式(13.7.10)则为

$$u^{(m+1)}=(I-\omega_{m+1}\hat{A})u^{(m)}+\omega_{m+1}\hat{b} \tag{13.7.14}$$

(13.7.10)与(13.7.13)相减，得到误差关系式

$$e^{(m+1)}=(I-\omega_{m+1}\hat{A})e^{(m)},\ m=0,\ 1,\ \cdots$$

由此递推，得到

$$e^{(m)}=G_m e^{(0)},\ G_m=\prod_{k=1}^{m}(I-\omega_k\hat{A}) \tag{13.7.15}$$

G_m 可以看成是迭代了 m 次时的收束阵，显然，它的本征值 $\eta_1,\ \cdots,\ \eta_n$ 与 $\hat{A}=D^{-1}A$ 的本征值 $\xi_1,\ \cdots,\ \xi_n$ 之间有对应关系

$$\eta_i=\prod_{k=1}^{m}(1-\omega_k\xi_i),\quad i=1,\ \cdots,\ n \tag{13.7.16}$$

命

$$P_m(\xi)=\prod_{k=1}^{m}(1-\omega_k\xi) \tag{13.7.17}$$

这是变元 ξ 的 m 次多项式，其零点为 $\omega_1^{-1},\ \omega_2^{-1},\ \cdots,\ \omega_m^{-1}$ 并且有

$$P_m(0)=1 \tag{13.7.18}$$

$$G_m=P_m(\hat{A}) \tag{13.7.19}$$

固定一个次数 m，以此 m 为周期，循环重复使用参数

$$\omega_1,\ \omega_2,\ \cdots,\ \omega_m,\ \omega_{m+1}=\omega_1,\ \omega_{m+2}=\omega_2,\ \cdots,\ \omega_{2m}=\omega_m,\ \omega_{2m+1}=\omega_1,\ \cdots$$

问题在于定出参数 $\omega_1,\ \cdots,\ \omega_m$ 使得迭代一个周期（m 次）后误差缩减最多，也就是使得 G_m 的本征值模量尽可能地小。

约定将 $\hat{A}$ 的本征值按小大次序排列

$$0<\xi_1\leqslant\xi_2\leqslant\cdots\leqslant\xi_{n-1}\leqslant\xi_n$$

也就是记最小及最大本征值为

$$\xi_{\min}=\xi_1,\quad \xi_{\max}=\xi_n \tag{13.7.20}$$

$\hat{A}$ 的本征值局限于区间 $\xi_1\leqslant\xi\leqslant\xi_n$ 内。因此 G_m 的本征值的最大模即谱半径 η_m 必满足

$$\eta_m=\rho(G_m)=\max_{1\leqslant i\leqslant n}|\eta_i|=\max_{1\leqslant i\leqslant n}\left|\prod_{k=1}^{m}(1-\omega_k\xi_i)\right|\leqslant\max_{\xi_1\leqslant\xi\leqslant\xi_n}\left|\prod_{k=1}^{m}(1-\omega_k\xi)\right| \tag{13.7.21}$$

于是不妨稍退一步，把问题转化为定出参数 $\omega_1,\ \cdots,\ \omega_m$ 使得

$$\max_{\xi_1\leqslant\xi\leqslant\xi_n}\left|\prod_{k=1}^{m}(1-\omega_k\xi)\right|=\text{极小} \tag{13.7.22}$$

注意 (13.7.17) 是满足条件 $P_m(0)=1$ 的 m 次多项式 $P_m(\xi)$ 的一般形式，因此问题就成为在条件 $P_m(0)=1$ 之下定出 m 次多项式 $P_m(\xi)$ 使得

$$\max_{\xi_1\leqslant\xi\leqslant\xi_n}|P_m(\xi)|=\text{极小} \tag{13.7.23}$$

为了方便，作线性变换把区间 $[\xi_1,\ \xi_2]$ 变成 $[-1,\ 1]$；为此取

$$x=\alpha-\beta\xi,\quad \xi=\beta^{-1}(\alpha-x) \tag{13.7.24}$$

此处

$$\alpha=\frac{\xi_n+\xi_1}{\xi_n-\xi_1}>1,\quad \beta=\frac{2}{\xi_n-\xi_1}>0 \tag{13.7.25}$$

于是

$$\xi_1\leqslant\xi\leqslant\xi_n\sim-1\leqslant x\leqslant-1$$

$$-\infty<\xi<\infty\sim-\infty<x<\infty$$

$$\xi=\xi_1,\ \xi_n\sim x=1,\ -1$$

$$\xi=0\sim x=\alpha=\frac{\xi_n+\xi_1}{\xi_n-\xi_1}>1$$

并且 ξ 的 m 次多项式 $P_m(\xi)$ 变为 x 的 m 次多项式 $Q_m(x)$

$$Q_m(d-\beta\xi)\equiv P_m(\xi),\quad P_m(\beta^{-1}(\alpha-x))\equiv Q_m(x) \tag{13.7.26}$$

于是上列问题(13.7.18)，(13.7.23)等价于在条件

$$Q_m(\alpha)=1 \quad (\alpha>1) \tag{13.7.27}$$

下定 m 次多项式 $Q_m(x)$ 使得

$$\max_{-1\leqslant x\leqslant 1}|Q_m(x)|=\text{极小} \tag{13.7.28}$$

这是一个经典的极小化问题，熟知它的解可以通过所谓 m 次切比雪夫多项式来表达。为了这个目的先将有关切氏多项式的一些基本性质罗列如下。

1．递推关系——作为 m 次切氏多项式 $T_m(x)$ 的定义。

$$T_0(x)=1,\ T_1(x)=x,\ T_{m+1}(x)=2xT_m(x)-T_{m-1}(x),\ m=1,\ 2,\ \cdots \tag{13.7.29}$$

2．对称或反对称性：

$$T_m(-x)\equiv(-)^m T_m(x) \tag{13.7.30}$$

3．在区间 $|x|\leqslant 1$ 上的表达式和最大模：

$$T_m(x)=\cos m\theta,\ \theta=\arccos x,\ 0\leqslant\theta\leqslant\pi,\ -1\leqslant x\leqslant 1 \tag{13.7.31}$$

$$\max_{-1\leqslant x\leqslant 1}|T_m(x)|=1=T_m(1)=|T_m(-1)| \tag{13.7.32}$$

4．在区间 $|x|\leqslant 1$ 外的表达式

$$T_m(x)=\cosh m\sigma=\frac{1}{2}[(x+\sqrt{x^2-1})^m+(x-\sqrt{x^2-1})^m],\ x\geqslant 1 \tag{13.7.33}$$

$$\sigma=\operatorname{arc\,cosh} x=\ln(x+\sqrt{x^2-1})$$

当 $x>1$ 时 $T_m(x)>1$ 并单调增至 ∞(当 $x\to\infty$)

5．零点——全部在 $(-1,\ 1)$ 之内

$$T_m(x_{m,k})=0,\ x_{m,k}=\cos(2k-1)\pi/2m,\ k=1,\ \cdots,\ m \tag{13.7.34}$$

6．相对极值点——全部在 $(-1,\ 1)$ 之内

$$T'_m(x'_{m,k})=0,\ T_m(x'_{m,k})=(-1)^k,\ x'_{m,k}=\cos k\pi/m,\ k=1,\ \cdots,\ m-1 \tag{13.7.35}$$

7．极值性质：对于任给的 $\alpha>1$，在所有满足 $Q_m(\alpha)=1$ 的 m 次多项式 $Q_m(x)$ 之中，是多项式

$$Q_m(x)=T_m(x)/T_m(\alpha) \tag{13.7.36}$$

使得

$$\max_{-1\leqslant x\leqslant 1}|Q_m(x)|=\text{极小}$$

并且所达到的极小模量就是

$$\max_{-1\leqslant x\leqslant 1}|Q_m(x)|=\max_{-1\leqslant x\leqslant 1}|T_m(x)/T_m(\alpha)|=1/T_m(\alpha) \tag{13.7.37}$$

正是这个极值性质提供了极小化问题(13.7.27～28)的解为(13.7.36)。通过线性变换(13.7.24)就得原问题(13.7.18)、(13.7.23)的解

$$P_m(\xi)=T_m(\alpha-\beta\xi)/T(\alpha),\ \alpha=\frac{\xi_n+\xi_1}{\xi_n-\xi_1}>1,\ \beta=\frac{2}{\xi_n-\xi_1}>0 \tag{13.7.38}$$

并且所达到的极小模量为

$$\max_{\xi_1<\xi<\xi_n}|P_m(\xi)|=1/T_m(\alpha) \tag{13.7.39}$$

与此同时，$T_m(x)$的零点(13.7.34)变为$P_m(\xi)$的零点

$$P_m(\alpha_k)=0,\ \alpha_k=\beta^{-1}[\alpha-\cos(2k-1)\pi/2m],\ k=1,\ \cdots,\ m \tag{13.7.40}$$

它们都在区间$(\xi_1,\ \xi_m)$之内。考虑到条件$P_m(0)=1$,

$$P_m(\xi)=\prod_{k=1}^{m}(1-\alpha_k^{-1}\xi) \tag{13.7.41}$$

与(13.7.17)相比较，便可以得到迭代参数，比如说

$$\omega_k=\alpha_k^{-1},\ k=1,\ \cdots,\ m \tag{13.7.42}$$

每一个参数ω_k对应于$P_m(\xi)$的一个零点。显然也可以将零点的次序任意排列，可以取

$$\omega_k=(\alpha_{\sigma_k})^{-1},\ k=1,\ \cdots,\ m \tag{13.7.43}$$

这里$\{\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_m\}$是$\{1,\ 2,\ \cdots,\ m\}$的一个任意排列。顺序的选择涉及计算的稳定性，详见13.7.4节。

由于(13.7.32)

$$\max_{\xi_1<\xi<\xi_n}|P_m(\xi)|=|P_m(\xi_1)|=|P_m(\xi_n)|=\max_{1<i<n}|P_m(\xi_i)|=\rho(G_m) \tag{13.7.44}$$

因此

$$\eta_m=\rho(G_m)=1/T(\alpha)<1 \tag{13.7.45}$$

迭代次数取为周期m的整数倍时方法收敛。由于这个方法是与切氏多项式相联系的，因此也叫做切氏迭代法或切氏加速法。

在这个方法中，优选迭代参数$\omega_1,\ \cdots,\ \omega_m$取决于两个参数即$\hat{\boldsymbol{A}}$的最小和最大本征值$\xi_1$, ξ_n。对于模型问题或者对于系数阵具有“性质A”的情况(13.5.4节)矩阵$\boldsymbol{G}$的本征值正负成对出现，$\hat{\boldsymbol{A}}=\boldsymbol{I}-\boldsymbol{G}$, 于是

$$\xi_1=\xi_{\min}=1-\mu,\ \xi_n=\xi_{\max}=1+\mu,\ \mu=\rho(\boldsymbol{G}) \tag{13.7.46}$$

因此为了优选，只用到一个参数μ，与超松弛法相同。这时

$$\left.\begin{aligned}
&\alpha=\frac{\xi_n+\xi_1}{\xi_n-\xi_1}=\frac{1}{\mu},\ \beta=\frac{2}{\xi_n-\xi_1}=\frac{1}{\mu}=\alpha\\
&P_m(\xi)=T_m\left(\frac{1-\xi}{\mu}\right)\Big/T_m\left(\frac{1}{\mu}\right)\\
&\eta_m=\rho(\boldsymbol{G}_m)=1/T\left(\frac{1}{\mu}\right)\\
&\alpha_k=1+\mu\cos(2k-1)\pi/2m,\ k=1,\ \cdots,\ m\\
&\omega_k=(\alpha_{\sigma_k})^{-1},\ k=1,\ \cdots,\ m
\end{aligned}\right\} \tag{13.7.47}$$

当取周期$m=1$时，$T_1(x)=x, P_m(\xi)=1-\xi,\ \alpha_1=1, \omega_1=1$。这就是普通的简单迭代法。

13.7.2 平均收敛速度

在变参数法中，阵$\boldsymbol{G}_m$的谱半径$\eta_m=\rho(\boldsymbol{G}_m)$可以迭代$m$次累计的收束因子，分摊到每次迭代则相当于收束因子$(\eta_m)^{\frac{1}{m}}$。因此不妨规定所谓平均收敛速度

$$R_m=-\ln(\eta_m)^{\frac{1}{m}}=\frac{1}{m}(-\ln\eta_m) \tag{13.7.48}$$

来估量收敛的快慢。如果各步迭代取同一常参数，则平均收敛速度与以前(13.5.2节)中规

定的渐近收敛速度 R 是一致的。这是因为，如果每次迭代的谱半径为 η，则迭代 m 次的谱半径为 $\eta_m=\eta^m$，分摊到每次则 $(\eta^m)^{\frac{1}{m}}=\eta$，所以 $R_m=R$。

我们将以模型问题为例来估算切氏迭代法的平均收敛速度，并与简单迭代和优选超松弛法相比较。为此，引用一些有关的参数

$$p=p(\boldsymbol{D}^{-1}\boldsymbol{A})=\frac{\xi_n}{\xi_1}=\frac{1+\mu}{1-\mu}=\operatorname{ctg}^2\pi/2N\approx 4N^2/\pi^2,\ \sqrt{p}\approx 2N/\pi \tag{13.7.49}$$

$$\mu=\rho(\boldsymbol{G}),\quad R=-\ln\mu \tag{13.7.50}$$

$$\eta=\rho(\boldsymbol{H}_{\omega*}),\quad R^*=-\ln\eta \tag{13.7.51}$$

这些参数只有一个是独立的，它们可以互相表达，例如

$$\eta=\frac{1-\sqrt{1-\mu^2}}{1+\sqrt{1-\mu^2}}=\left(\frac{\sqrt{p}-1}{\sqrt{p}+1}\right)^2\approx 1-\frac{4}{\sqrt{p}},\ R^*\approx\frac{4}{\sqrt{p}} \tag{13.7.52}$$

$$\mu=\frac{2\eta^{\frac{1}{2}}}{1+\eta}=\frac{p-1}{p+1}\approx 1-\frac{2}{p},\ R\approx\frac{2}{p} \tag{13.7.53}$$

根据(13.7.47)，(13.7.33)，对切氏加速法有

$$\eta_m=1/T_m\left(\frac{1}{\mu}\right)=\left\{\frac{1}{2}\left[\left(\frac{1}{\mu}+\sqrt{\frac{1}{\mu^2}-1}\right)^m+\left(\frac{1}{\mu}-\sqrt{\frac{1}{\mu^2}-1}\right)^m\right]\right\}^{-1}$$

由此以及(13.7.53)，可以得到 η_m 的两种表达式

$$\eta_m=\frac{2\mu^m}{(1+\sqrt{1-\mu^2})^m+(1-\sqrt{1-\mu^2})^m} \tag{13.7.54}$$

$$\eta_m=\frac{2\eta^{\frac{m}{2}}}{1+\eta^m} \tag{13.7.55}$$

以下分三种情况来估计 η_m 和 R_m

(1) $m\ll\sqrt{p}$ 即 $m\sim 1$ 的情况

利用公式(13.7.54)，由于

$$\sqrt{1-\mu^2}=\sqrt{1-\left(\frac{p-1}{p+1}\right)^2}=\frac{2\sqrt{p}}{p+1}\approx\frac{2}{\sqrt{p}}$$

$$\begin{aligned}(1+\sqrt{1-\mu^2})^m+(1-\sqrt{1-\mu^2})^m&\approx\left(1+\frac{2}{\sqrt{p}}\right)^m+\left(1-\frac{2}{\sqrt{p}}\right)^m\\&\approx 1+\frac{4}{\sqrt{p}}+\frac{2m(m-1)}{p}+1-\frac{4}{\sqrt{p}}+\frac{2m(m-1)}{p}\\&=2\left[1+\frac{2m(m-1)}{p}\right]\end{aligned}$$

$$\eta_m\approx\mu^m\left[1-\frac{2m(m-1)}{p}\right]$$

$$-\frac{1}{m}\ln\eta_m=\frac{2}{p}-\frac{1}{m}\ln\left(1-\frac{2m(m-1)}{p}\right)\approx\frac{2}{p}+\frac{2(m-1)}{p}=\frac{2m}{p}$$

因此

$$R_m\approx\frac{2m}{p}\approx mR \tag{13.7.56}$$

即平均收敛速度约为简单迭代法的 m 倍。

(2) $m \gg \sqrt{p}$ 即 $m \sim \infty$ 的情况

利用公式(13.7.55),其中因子 $\frac{2}{1+\eta^m}$ 当 m 很大时趋于定限 2，于是

$$-\frac{1}{m}\ln\eta_m = -\frac{1}{m}\cdot\frac{m}{2}\ln\eta - \frac{1}{m}\ln\frac{2}{1+\eta^m} \approx -\frac{1}{2}\ln\eta$$

因此

$$R_m \approx \frac{2}{\sqrt{p}} \approx \frac{R^*}{2} \tag{13.7.57}$$

平均收敛速度约为优选超松弛法之半。

(3) $m \sim \sqrt{p}$ 即 m 与 $\sqrt{p}$ 相同量级的情况

利用公式(13.7.55)。命 $m = k\sqrt{p}$，$k \sim 1$，于是

$$\eta^m = \left(1-\frac{4}{\sqrt{p}}\right)^{k\sqrt{p}} = \left(1-\frac{4}{\sqrt{p}}\right)^{\frac{\sqrt{p}}{4}\cdot 4k} \approx e^{-4k}$$

$$-\frac{1}{m}\ln\eta_m \approx -\frac{1}{2}\ln\eta - \frac{1}{m}[\ln 2 - \ln(1+e^{-4k})]$$

$$\approx \frac{2}{\sqrt{p}}\left\{1 - \frac{\sqrt{p}}{2m}[\ln 2 - \ln(1+e^{-4k})]\right\}$$

$$R_m \approx \frac{2}{\sqrt{p}}c_k, \quad c_k = 1 - \frac{1}{2k}[\ln 2 - \ln(1+e^{-4k})] \tag{13.7.58}$$

c_k 不依赖于 p。可以算出具体数字得

$$\left.\begin{aligned} &m \approx \sqrt{p}: \ R_m \approx 0.66\frac{2}{\sqrt{p}} \\ &m \approx 2\sqrt{p}: \ R_m \approx 0.83\frac{2}{\sqrt{p}} \\ &m \approx 3\sqrt{p}: \ R_m \approx 0.89\frac{2}{\sqrt{p}} \\ &m \approx 4\sqrt{p}: \ R_m \approx 0.93\frac{2}{\sqrt{p}} \end{aligned}\right\} \tag{13.7.59}$$

由此可见，切氏迭代周期取 $m=\infty$ 时，达到最高的收敛速度 $R_\infty = \frac{1}{2}R^*$，相当于优选超松弛法的一半。关于如何实现 $m=\infty$ 的问题见 13.7.5 节。对于有穷的周期，m 应取得足够大，达到 $\sqrt{p}$ 的量级。比如，取 $m \approx 3\sqrt{p}$，对模型问题 $\sqrt{p} \approx 2N/\pi$，即 $m \approx 2N$，这时达到最高速度的 90%。

13.7.3 有关参数的试选方法

在切氏迭代法中，与超松弛法一样，要用到 $\mu = \rho(G)$ 来计算优选迭代参数。当 μ 值取得偏大或偏小时都导致收敛速度的降低。为了达到优选效果，需要对 μ 进行予估，可以在迭代过程中实验地定出 μ，方法与 § 13.6 相似，仅细节上有些不同。

一个简单的方法是在开始迭代时取常参数 $\omega_k \equiv 1$，也就是作简单迭代。与(13.6.8~9)，相仿，每步的余量或增量有下列关系

$$\boldsymbol{r}^{(k+1)} = \boldsymbol{AGA}^{-1}\boldsymbol{r}^{(k)} \tag{13.7.60}$$

$$\boldsymbol{d}^{(k+1)} = \boldsymbol{Gd}^{(k)} \tag{13.7.61}$$

AGA^{-1} 与 G 有相同的本征值。可以根据余量（或增量）利用幂法来定 G 的最大模本征值。比如说，每步取分量的最大模

$$r_k=\max_{1\leq i\leq n}|r_i^{(k)}| \tag{13.7.62}$$

作为余量的度量。由于 G 的最大模本征值为 μ, $-\mu$ 两个，而 G^2 的最大模本征值只有一个双重的 μ^2。因此 r_{k+1}/r_k 不收敛而 r_{k+2}/r_k 收敛。可以迭代到后一比值稳定而得到

$$r_{k+2}/r_k\to\mu^2 \tag{13.7.63}$$

对于增量也一样。这是一种"跳步"的幂法。反之，考虑到 G 的对称性，可以取增量的内积

$$d'_k=(d^{(k)},\ d^{(k)})=\sum_{i=1}^{n}(d_i^{(k)})^2 \tag{13.7.64}$$

作为一种度量。这时就无需跳步。事实上

$$d'_{k+1}=(d^{(k+1)},\ d^{(k+1)})=(Gd^{(k)},\ Gd^{(k)})=(G^TGd^{(k)},\ d^{(k)})=(G^2d^{(k)},\ d^{(k)})$$

因此

$$d'_{k+1}/d'_k\to\mu^2 \tag{13.7.65}$$

取得 μ 值后可以根据 $\sqrt{p}=\sqrt{\dfrac{1+\mu}{1-\mu}}$ 和 (13.7.59) 来估计应取的周期 m，然后根据 (13.7.47)，定出参数 ω_1, …, ω_m，转而用这些参数进行迭代。

顺便指出，切氏迭代的收敛范围是比较宽的。以上是在矩阵 $\hat{A}=D^{-1}A$ 的本征值都是正实数的基础上导出切氏迭代法的。设想把方法中用到的参数 α, β 视为形式参数，满足 $\alpha>1$, $\beta>0$。相应的切氏迭代一个周期的谱半径为

$$\eta_m=\rho(G_m)=\max_{1\leq i\leq n}|T_m(\alpha-\beta\xi)|/T_m(\alpha) \tag{13.7.66}$$

ξ_1, …, ξ_n 表示 $\hat{A}$ 的本征值。因此，即使 $\hat{A}$ 的本征值不全是正实数，或虽都是正实数而其最小最大本征值 ξ_1, ξ_n 与 α, β 之间不满足关系式 (14.7.25)，只要上式右端的值小于 1，迭代就是收敛的。当然只当 $\hat{A}$ 的本征值全为正实数并且满足(13.7.25)时方法才具有某种最优性。

13.7.4 不稳定性和稳化方法

一个严重的问题在于当周期 m 比较大时，按自然顺序(13.7.42)取参数 $\omega_1=\alpha_1^{-1}, \omega_2, \cdots, \omega_m=\alpha_m^{-1}$ 时计算不稳定。这时 ω_1^{-1}, ω_2^{-1}, …, ω_m^{-1} 在 ξ 轴上是自右至左即从 $\hat{A}$ 的最大本征值 $\xi_n\approx1+\mu$ 一端起逐步左移至最小本征值 $\xi_1\approx1-\mu$。在迭代初期，例如 $k=1$ 时

$$e^{(1)}=(I-\omega_1\hat{A})e^{(0)},\quad \omega_1\approx\frac{1}{1+\mu}$$

相应的误差曲线为

$$\eta\approx1-\frac{\xi}{1+\mu}\approx\begin{cases}0, & \xi\approx\xi_n\\ \dfrac{2\mu}{1+\mu}, & \xi\approx\xi_1\end{cases}$$

这表示对高频误差(相当于 $\hat{A}$ 的大本征值的误差分量)大大压缩到近于 0，而对低频误差(相当 $\hat{A}$ 的小本征值)则有少许压缩。反之，在周期终了时，例如 $k=m$ 时

$$e^{(m)}=(I-\omega_m\hat{A})e^{(m-1)},\quad \omega_m\approx\frac{1}{1-\mu}$$

相应的误差曲线是

$$\eta\approx1-\frac{\xi}{1-\mu}\approx\begin{cases}0,\ \xi\approx\xi_1\\ -\dfrac{2\mu}{1-\mu}\approx-p,\ \xi\approx\xi_n\end{cases}$$

这时对低频误差大大压缩近于0而对高频误差则大大地放大。

这样，在一个周期结束时，虽然根据(13.7.45)理论上误差被一致压缩，但是在迭代的过程中每步都有舍入误差，基本上分摊到各个本征分量上，到了后半周期，高频误差连续地作大量的放大，造成误差的恶性积累。因此切氏迭代法按照自然顺序(13.7.42)实质上是不可行的。必须采取稳化的措施。

稳定化的一种方法是把单层迭代改造双层迭代。迄今为止，松弛法和简单迭代法都是所谓单层的，即计算 $u^{(k+1)}$ 时只用到一个层次 $u^{(k)}$，而在双层格式中则要用到 $u^{(k)}$ 和 $u^{(k-1)}$ 两个层次。在双层形式下不仅可以实现稳化，还可实现无穷大周期 $m=\infty$，缺点是需要存储两片场，详见 13.7.5 节。这里先介绍另一种比较简单的稳化法，只是将切氏多项式的零点 $\alpha_1, \cdots, \alpha_m$ 的次序重新排列，其它一切照旧，保持了单层的优点。

办法是定出一个适当的排列

$$\{1,\ 2,\ \cdots,\ m\}\rightarrow\{\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_m\}$$

也就是

$$\{\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_m\}\rightarrow\{\alpha_{\sigma_1},\ \alpha_{\sigma_2},\ \cdots,\ \alpha_{\sigma_m}\}$$

使得依次取参数如(4.43)即

$$\omega_1=(\alpha_{\sigma_1})^{-1},\ \omega_2=(\alpha_{\sigma_2})^{-1},\ \cdots,\ \omega_m=(\alpha_{\sigma_m})^{-1}$$

时每步迭代都保证稳定。

显然可见，不论采取什么顺序，迭代一个周期后的误差多项式是相同的

$$P_m(\xi)\equiv\prod_{k=1}^{m}(1-\alpha_k^{-1}\xi)\equiv\prod_{k=1}^{m}(1-\alpha_{\sigma_k}^{-1}\xi)$$

对于迭代的每一次 $k=1, \cdots, m$　P_m 可以分解为两个因子

$$P_m(\xi)=Q_{m,k}(\xi)R_{m,k}(\xi)$$

$$Q_{m,k}(\xi)=\prod_{i=1}^{k}(1-\alpha_{\sigma_i}^{-1}\xi),\ R_{m,k}(\xi)=\prod_{i=k+1}^{m}(1-\alpha_{\sigma_i}^{-1}\xi)$$

$Q_{m,k}$ 表示前 k 次的误差多项式，$R_{m,k}$ 表示后 $m-k$ 次的误差多项式。

$Q_{m,k}$，$R_{m,k}$ 分别表示前 k 次和后 $m-k$ 次的误差多项式。为了保证全过程的稳定性，应该要求

$$\max_{\xi_1<\xi<\xi_n}|Q_{m,k}(\xi)|,\ \max_{\xi_1<\xi<\xi_n}|R_{m,k}(\xi)|$$

对于 $k=1, 2, \cdots, m$ 以及 $m=1, 2, \cdots$ 为一致有界。

这种保证稳定性的排列 σ_i 是可以找到的。当周期 m 为 2 的整次幂即 $m=2^s$ ($s\geqslant0$ 整数) 时特别简单，可以递推地产生[5]。

对于 $m=2^s$ 的排列记为 $\{\sigma_1^{(s)}, \cdots, \sigma_{2^s}^{(s)}\}$。取 $\{\sigma_1^{(0)}\}=\{1\}$，依次把 $\{\sigma_1^{(s-1)}, \cdots, \sigma_{2^{s-1}}^{(s-1)}\}$ 的第 i 项 $\sigma_i^{(s-1)}$ 代以两项 $\sigma_i^{(s-1)}$, $2^s+1-\sigma_i^{(s-1)}$ 即得排列 $\{\sigma_1^{(s)}, \cdots, \sigma_{2^s}^{(s)}\}$。这就是说

$$\sigma_1^{(0)}=1$$

$$\sigma_{2i-1}^{(s)}=\sigma_i^{(s)},\ \sigma_{2i}^{(s)}=2^s+1-\sigma_i^{(s-1)},\ i=1,\ \cdots,\ 2^{s-1};\ s=1,\ 2,\ \cdots \tag{13.7.67}$$

注意这里每一对 σ_{2i-1}, σ_{2i} 对应于切氏多项式左右对称的两个零点的标号。当 $s=0\sim4$ 时排列的具体形式如下

$m=2^0=1,\quad \{1\}$

$m=2^1=2,\quad \{1,\ 2\}$

$m=2^2=4,\quad \{1,\ 4,\ 2,\ 3\}$

$m=2^3=8,\quad \{1,\ 8,\ 4,\ 5,\ 2,\ 7,\ 3,\ 6\}$

$m=2^4=16,\quad \{1,\ 16,\ 8,\ 9,\ 4,\ 13,\ 5,\ 12,\ 2,\ 15,\ 7,\ 10,\ 3,\ 14,\ 6,\ 11\}$

13.7.5 递推的切氏迭代法

在以上我们是事先规定周期 m，通过 m 个零点来构成切氏多项式和相应的单层迭代。事实上切氏多项式还可以通过"三项"递推公式来构成，从而得到双层的即联系到"三项" $\boldsymbol{u}^{(m+1)}$, $\boldsymbol{u}^{(m)}$, $\boldsymbol{u}^{(m-1)}$ 的迭代格式。这种迭代法是稳定的，并且无须预定周期，相当取周期 $m=\infty$，因此达到最优的收敛速度 R_∞，为优选超松弛法的一半。对于模型问题或一般的五点差分格式则还可以进一步改造为单层的迭代格式，收敛速度提高到与优选超松弛法相等，但比后者改善了初期收敛性，这就是 13.5.5 节中的变参数超松弛法。

我们从误差公式(13.7.15)，(13.7.19)出发，

$$\boldsymbol{e}^{(m)}=P_m(\hat{\boldsymbol{A}})\boldsymbol{e}^{(0)},\quad m=1,\ 2,\ \cdots \tag{13.7.68}$$

$$P_m(\xi)=\frac{1}{\tau_m}T_m(\alpha-\beta\xi) \tag{13.7.69}$$

$$\tau_m=T_m(\alpha) \tag{13.7.70}$$

$$\alpha=\frac{\xi_n+\xi_1}{\xi_n-\xi_1},\quad \beta=\frac{2}{\xi_n-\xi_1} \tag{13.7.71}$$

利用 T_m 的"三项"递推式

$$T_0(x)=1,\ T_1(x)=x,\ T_{m+1}(x)=2xT_m(x)-T_{m-1}(x)$$

可以导出 τ_m, $P_m(\xi)$, $P_m(\hat{A})$ 的"三项"递推式

$$\tau_0=1,\ \tau_1=\alpha,\ \tau_{m+1}=2\alpha\tau_m-\tau_{m-1} \tag{13.7.72}$$

$$\begin{aligned}P_0(\xi)=1,\ P_1(\xi)=\frac{1}{\tau_1}(\alpha-\beta\xi),\ \tau_{m+1}P_{m+1}(\xi)\\ =2\tau_m(\alpha-\beta\xi)P_m(\xi)-\tau_{m-1}P_{m-1}(\xi)\end{aligned} \tag{13.7.73}$$

$$\begin{aligned}P_0(\hat{\boldsymbol{A}})=\boldsymbol{I},\ P_1(\hat{\boldsymbol{A}})=\frac{1}{\tau_1}(\alpha\boldsymbol{I}-\beta\hat{\boldsymbol{A}}),\ \tau_{m+1}P_{m+1}(\hat{\boldsymbol{A}})\\ =2\tau_m(\alpha\boldsymbol{I}-\beta\hat{\boldsymbol{A}})P_m(\hat{\boldsymbol{A}})-\tau_{m-1}P_{m-1}(\hat{\boldsymbol{A}})\end{aligned} \tag{13.7.74}$$

对最后一式两端右乘以向量 $\boldsymbol{e}^{(0)}$，即得

$$\left.\begin{aligned}\boldsymbol{e}^{(1)}=\frac{1}{\tau_1}(\alpha\boldsymbol{I}-\beta\hat{\boldsymbol{A}})\boldsymbol{e}^{(0)}\\ \tau_{m+1}\boldsymbol{e}^{(m+1)}=2\tau_m(\alpha\boldsymbol{I}-\beta\hat{\boldsymbol{A}})\boldsymbol{e}^{(m)}-\tau_{m-1}\boldsymbol{e}^{(m-1)}\end{aligned}\right\} \tag{13.7.75}$$

命 $\boldsymbol{u}$ 为代数方程组的真解

$$\hat{\boldsymbol{A}}\boldsymbol{u}=\hat{\boldsymbol{b}},\quad \boldsymbol{e}^{(m)}=\boldsymbol{u}^{(m)}-\boldsymbol{u} \tag{13.7.76}$$

以此代入(13.7.75)并利用 τ_m 的递推式(13.7.72)可得

$$\begin{cases} \boldsymbol{u}^{(1)}=\boldsymbol{u}^{(0)}+\dfrac{\beta}{\alpha}(\hat{\boldsymbol{b}}-\hat{\boldsymbol{A}}\boldsymbol{u}^{(0)}) \\ \boldsymbol{u}^{(m+1)}=\dfrac{1}{\tau_{m+1}}[2\alpha\tau_m\boldsymbol{u}^{(m)}+(\tau_{m+1}-2\alpha\tau_m)\boldsymbol{u}^{(m-1)}+2\beta\tau_m(\hat{\boldsymbol{b}}-\hat{\boldsymbol{A}}\boldsymbol{u}^{(m)})], \\ \qquad m=1,\ 2,\ \cdots \end{cases} \tag{13.7.77}$$

命

$$\omega_1=1,\quad \omega_{m+1}=2\alpha\tau_m/\tau_{m+1},\quad m=1,\ 2,\ \cdots \tag{13.7.78}$$

于是上式可以统一表为

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m-1)}+\omega_{m+1}\left[\frac{\beta}{\alpha}(\hat{\boldsymbol{b}}-\hat{\boldsymbol{A}}\boldsymbol{u}^{(m)})+\boldsymbol{u}^{(m)}-\boldsymbol{u}^{(m-1)}\right],\quad m=0,\ 1,\ \cdots \tag{13.7.79}$$

这里初始场 $\boldsymbol{u}^{(0)}$ 可以任取并约定 $\boldsymbol{u}^{(-1)}=0$。上式也可表为“增量”的形式，可能更方便些

$$\begin{cases} \boldsymbol{d}^{(m+1)}=(\omega_{m+1}-1)\boldsymbol{d}^{(m)}+\omega_{m+1}\cdot\dfrac{\beta}{\alpha}(\hat{\boldsymbol{b}}-\hat{\boldsymbol{A}}\boldsymbol{u}^{(m)}) \\ \boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m)}+\boldsymbol{d}^{(m+1)},\quad m=0,\ 1,\ \cdots \end{cases} \tag{13.7.80}$$

这里约定初始增量场 $\boldsymbol{d}^{(0)}=\boldsymbol{u}^{(0)}$，$\boldsymbol{u}^{(0)}$ 可以任取。

参数 ω_m 可以直接用递推式产生

$$\omega_1=1,\ \omega_2=\frac{2\alpha^2}{2\alpha^2-1},\quad \omega_{m+1}=\frac{4\alpha^2}{4\alpha^2-\omega_m},\quad m=2,\ 3,\ \cdots \tag{13.7.81}$$

这是因为，根据(13.7.78)，(13.7.72)，

$$\omega_1=\alpha\tau_0/\tau_1=1$$

$$\omega_2=2\alpha\tau_1/\tau_2=2\alpha^2/(2\alpha\tau_1-\tau_2)=2\alpha^2/(2\alpha^2-1)$$

而方程组

$$\omega_{m+1}\tau_{m+1}-2\alpha\tau_m=0$$

$$\omega_m\tau_m-2\alpha\tau_{m-1}=0$$

$$\tau_{m+1}-2\alpha\tau_m+\tau_{m-1}=0$$

有非零解 τ_{m+1}，τ_m，τ_{m-1}，故其系数行列式为零，即

$$\omega_{m+1}\omega_m+4\alpha^2-4\alpha^2\omega_{m+1}=0$$

这就是(13.7.81)。

对模型问题以及系数阵具有所谓“性质 A”(包括一般的五点差分格式)，

$$\xi_1=1-\mu,\ \xi_n=1+\mu,\ \alpha=\beta=\frac{1}{\mu},\ \mu=\beta(\boldsymbol{G}),\ \boldsymbol{G}=\boldsymbol{I}-\hat{\boldsymbol{A}} \tag{13.7.82}$$

于是(13.7.79)，(13.7.80)分别简化为:

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m-1)}+\omega_{m+1}[\hat{\boldsymbol{b}}-\hat{\boldsymbol{A}}\boldsymbol{u}^{(m)}+\boldsymbol{u}^{(m)}-\boldsymbol{u}^{(m-1)}],\quad m=0,\ 1,\ \cdots \tag{13.7.83}$$

即

$$\boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m-1)}+\omega_{m+1}[\hat{\boldsymbol{b}}+\boldsymbol{G}\boldsymbol{u}^{(m)}-\boldsymbol{u}^{(m-1)}],\quad m=0,\ 1,\ \cdots \tag{13.7.84}$$

$$\begin{cases} \boldsymbol{d}^{(m+1)}=(\omega_{m+1}-1)\boldsymbol{d}^{(m)}+\omega_{m+1}(\hat{\boldsymbol{b}}-\hat{\boldsymbol{A}}\boldsymbol{u}^{(m)}) \\ \boldsymbol{u}^{(m+1)}=\boldsymbol{u}^{(m)}+\boldsymbol{d}^{(m+1)},\quad m=0,\ 1,\ \cdots \end{cases} \tag{13.7.85}$$

$$\omega_1=1,\quad \omega_2=\frac{2}{2-\mu^2},\quad \omega_{m+1}=\frac{4}{4-\omega_m\mu^2},\quad m=2,\ 3,\ \cdots \tag{13.7.86}$$

迭代式(13.7.79)或(13.7.80)是双层的，是切氏迭代法的一种变形，它们实现了切氏多项式的误差规律

$$\boldsymbol{e}^{(m)}=P_m(\hat{\boldsymbol{A}})\boldsymbol{e}^{(0)}$$

$$\eta_m=\rho(P_m(\hat{\boldsymbol{A}}))=1/T_m(\alpha),\ \ m=1,\ 2,\ \cdots$$

设 $\hat{\boldsymbol{A}}$ 是对称阵(在模型问题确是如此)，则 $P_m(\hat{\boldsymbol{A}})$ 也对称阵。对称阵的欧氏模量等于其谱半径,因此

$$\|\boldsymbol{e}^{(m)}\|\leqslant\|P_m(\hat{\boldsymbol{A}})\|\cdot\|\boldsymbol{e}^{(0)}\| \tag{13.7.87}$$

$$\|P_m(\hat{\boldsymbol{A}})\|=\rho(P_m(\hat{\boldsymbol{A}}))=\eta_m=1/T_m(\alpha),\ \ m=1,\ 2,\ \cdots \tag{13.7.88}$$

根据(13.7.55), (13.7.52),

$$\|P_m(\boldsymbol{A})\|=\frac{2\eta^{\frac{m}{2}}}{1+\eta^m},\ \ \eta=\frac{1-\sqrt{1-\mu^2}}{1+\sqrt{1-\mu^2}}$$

这是m的单调递减函数,当 $m\to\infty$ 时趋于 0。因此，对于欧氏模量而言，误差是递降的，保证了方法的稳定性。注意这里 $m=0,\ 1,\ \cdots$ 可以无限增长而无需预定周期，实际上实现了无穷大的周期，因此可以发挥切氏迭代的最高效率。作为双层格式，对于未知向量 $\boldsymbol{u}$ 需要存储两片场,在有些问题中会成为负担,这是缺点。

对于模型问题或系数阵具有“性质 A”(包括一般的五点差分格式),在奇偶顺序下,差分方程组

$$\boldsymbol{A}\boldsymbol{u}=\boldsymbol{b}$$

的系数阵 $\boldsymbol{A}$ 表为特殊的形式(13.5.55)

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{D}_1 & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{D}_2\end{bmatrix},\qquad \boldsymbol{D}=\begin{bmatrix}\boldsymbol{D}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{D}_2\end{bmatrix} \tag{13.7.89}$$

$\boldsymbol{D}_1$, $\boldsymbol{D}_2$ 为对角阵，于是

$$\hat{\boldsymbol{A}}=\boldsymbol{D}^{-1}\boldsymbol{A}=\begin{bmatrix}\boldsymbol{I} & \hat{\boldsymbol{A}}_{12}\\ \hat{\boldsymbol{A}}_{21} & \boldsymbol{I}\end{bmatrix},\qquad \hat{\boldsymbol{A}}_{12}=\boldsymbol{D}^{-1}\boldsymbol{A}_{12},\hat{\boldsymbol{A}}_{21}=\boldsymbol{D}^{-1}\boldsymbol{A}_{21} \tag{13.7.90}$$

$$\boldsymbol{G}=\boldsymbol{I}-\hat{\boldsymbol{A}}=\begin{bmatrix}0 & \boldsymbol{G}_{12}\\ \boldsymbol{G}_{21} & 0\end{bmatrix},\qquad \boldsymbol{G}_{12}=-\hat{\boldsymbol{A}}_{12},\boldsymbol{G}_{21}=-\hat{\boldsymbol{A}}_{21} \tag{13.7.91}$$

相应地把向量 $\boldsymbol{u}$, $\boldsymbol{b}$, $\hat{\boldsymbol{b}}=\boldsymbol{D}^{-1}\boldsymbol{b}$ 分解为奇、偶两段

$$\boldsymbol{u}=\begin{bmatrix}u_1\\ u_2\end{bmatrix},\qquad \boldsymbol{b}=\begin{bmatrix}b_1\\ b_2\end{bmatrix},\qquad \hat{\boldsymbol{b}}=\boldsymbol{D}^{-1}\boldsymbol{b}=\begin{bmatrix}\hat{\boldsymbol{b}}_1\\ \hat{\boldsymbol{b}}_2\end{bmatrix}=\begin{bmatrix}\boldsymbol{D}_1^{-1} & \boldsymbol{b}_1\\ \boldsymbol{D}_2^{-1} & \boldsymbol{b}_2\end{bmatrix} \tag{13.7.92}$$

于是迭代格式(13.7.14)可以表为

$$\begin{cases}\text{奇点:}\ \ \boldsymbol{u}_1^{(m+1)}=\boldsymbol{u}_1^{(m-1)}+\omega_{m+1}(\hat{\boldsymbol{b}}_1-\boldsymbol{G}_{12}\boldsymbol{u}_2^{(m)}-\boldsymbol{u}_1^{(m-1)})\\ \text{偶点:}\ \ \boldsymbol{u}_2^{(m+1)}=\boldsymbol{u}_2^{(m-1)}+\omega_{m+1}(\hat{\boldsymbol{b}}_2-\boldsymbol{G}_{21}\boldsymbol{u}_1^{(m)}-\boldsymbol{u}_2^{(m-1)})\end{cases} \tag{13.7.93}$$

为了弄清楚算法的特点,写下奇、偶点以及奇、偶次的迭代公式

$$\begin{aligned}&\text{奇次, 奇点:}\ \ \boldsymbol{u}_1^{(2m+1)}=\boldsymbol{u}_1^{(2m-1)}+\omega_{2m+1}(\hat{\boldsymbol{b}}_1-\boldsymbol{G}_{12}\boldsymbol{u}_2^{(2m)}-\boldsymbol{u}_1^{(2m-1)})\\ &\text{奇次, 偶点:}\ \ \boldsymbol{u}_2^{(2m+1)}=\boldsymbol{u}_2^{(2m-1)}+\omega_{2m+1}(\hat{\boldsymbol{b}}_2-\boldsymbol{G}_{21}\boldsymbol{u}_1^{(2m)}-\boldsymbol{u}_2^{(2m-1)})\\ &\text{偶次, 奇点:}\ \ \boldsymbol{u}_1^{(2m+2)}=\boldsymbol{u}_1^{(2m)}+\omega_{2m+2}(\hat{\boldsymbol{b}}_1-\boldsymbol{G}_{12}\boldsymbol{u}_2^{(2m+1)}-\boldsymbol{u}_1^{(2m)})\\ &\text{偶次, 偶点:}\ \ \boldsymbol{u}_2^{(2m+2)}=\boldsymbol{u}_2^{(2m)}+\omega_{2m+2}(\hat{\boldsymbol{b}}_2-\boldsymbol{G}_{21}\boldsymbol{u}_1^{(2m+1)}-\boldsymbol{u}_2^{(2m)})\end{aligned}$$

注意在奇次奇点或偶次偶点工作时只用到奇次奇点和偶次偶点的值，因此奇次偶点和偶次奇点的迭代可以略去不算而不影响迭代的进程,在奇点上只保留奇次值,偶点上只保留偶次值。因此得到

$$\begin{cases} \text{奇次, 奇点:} & u_1^{(2m+1)} = u_1^{(2m-1)} + \omega_{2m+1}(\hat{b}_1 - G_{12}u_2^{(2m)} - u_1^{(2m-1)}) \\ \text{偶次, 偶点:} & u_2^{(2m+2)} = u_2^{(2m)} + \omega_{2m+2}(\hat{b}_2 - G_{21}u_1^{(2m+1)} - u_2^{(2m)}) \end{cases} \quad (13.7.94)$$

节约了一半存储量，又节约了一半工作量，相当于收敛加快一倍。这里把普通意义下的全场一次迭代看作半场两次迭代，因此在普通意义下的 m 次迭代相当于 $2m$ 次切比雪夫迭代。这个方法就是 13.5.5 节所述的一种变参数超松弛法 (13.5.78～79)，其渐近收敛速度同于优选超松弛，但比后者改善了初期收敛性(13.5.6 节)。

参 考 资 料

[1] 福雪斯，华沙，《偏微分方程的差分方法》，科学出版社，1965。

[2] 瓦尔格，《矩阵迭代分析》，上海科学技术出版社，1966。

[3] Hockney, R. W. "The potential calculation and some applications", 载于 Methods in Computational Physics, Vol. 9, Plasma Physics, Academic Press, 1970, pp. 135–211.

[4] 加藤敏夫，《变分法及其应用》，上海科学技术出版社，1961。

[5] Лебедев В. И., Финогенов С. А., "О порядке выбора итерационных параметров в чебышевском циклическом итерационном методе", Журнал Вычислительной Математики и Математической Физики 11: 2 (1971) pp. 425–438.

[6] 黄鸿慈，《关于椭圆型方程 Neumann 问题的数值解法》，应用数学与计算数学 1:2 (1964), 121-130 页。

第十四章　有限元方法

有限元方法是椭圆型方程问题的一类数值解法。它的基础分两个方面：一是变分原理，二是剖分插值。从第一方面看，它是传统的能量法即李兹-加辽金方法的一种变形。从第二方面看则它是差分方法即格网法的一种变形。这是两类方法相结合取长补短而进一步发展了的结果，它具有很广泛的适应性，特别适合于几何、物理条件比较复杂的问题，而且便于程序的标准化。§14.1 介绍与椭圆方程相等价的变分原理。§14.2 介绍剖分插值，重点是三角剖分和相应的线性插值。§14.3 以典型的二阶椭圆方程问题为例说明有限元离散化的全过程。§14.4 介绍有限元法的一些应用。至于方法对四阶椭圆方程的推广以及对于众多物理、技术领域的应用则可参考专门的著作。

§14.1　变分原理

14.1.1　椭圆方程的变分原理

一般的椭圆型方程边值问题都有适当的变分原理与之等价。作为典型的例子，取平面域 Ω 上的二阶变系数椭圆型方程

$$\Omega:\quad -\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)=f \tag{14.1.1}$$

这里 $\beta=\beta(x, y)>0$, $f=f(x, y)$ 都是予给的分布。物理上众多的平衡态和定常态问题都归结这个典型的方程，或其简化了的或推广了的形式，例如弹性膜的平衡，弹性柱体的扭转、定常态的热传导或扩散，静电、磁场，不可压缩无旋流，定常渗流，定常亚声速流等等。

由于方程(14.1.1)对于导数是二阶的，为了保证唯一定解在边界 $\partial\Omega$ 上要给定一个条件。边界条件通常有三种类型(见第十三章 §13.1)：

第一类：$u=\bar{u}$

第二类：$\beta\dfrac{\partial u}{\partial \nu}=q$

第三类：$\beta\dfrac{\partial u}{\partial \nu}+\eta u=q$

这里 $\bar{u}$, q, η 为给定在边界上的分布，β 就是 $\beta(x, y)$ 在边界上的值，$\beta>0$, $\boldsymbol{\nu}$ 为外法向，$\eta\geqslant 0$。在边界的不同区段上可以取不同类型的边界条件。由于第二类边界条件可以看作第三类当 $\eta\equiv 0$ 时的特例，故边界条件一般地可以表为

$$\Gamma_0:\quad u=\bar{u} \tag{14.1.2}$$

$$\Gamma_0':\quad \beta\frac{\partial u}{\partial \nu}+\eta u=q \tag{14.1.3}$$

Γ_0 及 Γ_0' 为 $\partial\Omega$ 上互补的两个部分，

$$\partial\Omega=\Gamma_0+\Gamma_0'$$

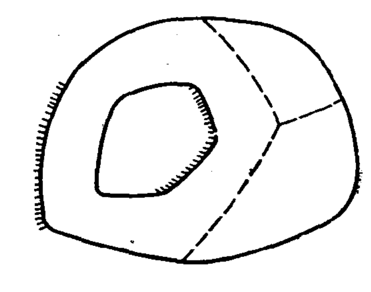


图　14.1

它们本身又可能分解为几个不相连结的区段。图 14.1 表示一个复连通域，边界上打毛的区段为 Γ_0。

对应于方程(14.1.1)和边界条件(14.1.3)可以构成所谓“能量积分”

$$J(u)=\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy+\int_{\Gamma_0'}\left[\frac{1}{2}\gamma u^2-qu\right]ds \quad (14.1.4)$$

任取一个函数 $u=u(x, y)$，有一个积分值 $J(u)$ 与之相应，因此 $J(u)$ 是“函数的函数”，可以叫做泛函。这里 J 二次地依赖于 u(的导数)，因此是一个二次泛函。

重要的事实在于：由所有满足边界条件

$$\Gamma_0:\ u=\bar{u}$$

的函数组成的函数类 S 中使得 J 达到极小值的那个函数即极值函数 $u=u(x, y)$ 必定在 Ω 内满足微分方程(14.1.1)而且在边界上除了在 Γ_0 上满足(14.1.2)以外还在 Γ_0' 上自动满足边界条件(14.1.3)；反之，满足(14.1.1～3)的函数 $u=u(x, y)$ 也必定是函数类 S 中使得 J 达到极小值的函数。这就是说变分问题

$$\begin{cases} J(u)=\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy+\int_{\Gamma_0'}\left[\frac{1}{2}\gamma u^2-qu\right]ds=\text{极小} & (14.1.5)\\ \Gamma_0:\ u=\bar{u} & (14.1.6)\end{cases}$$

等价于边值问题

$$\begin{cases} \Omega:\ -\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial y}\right)=f & (14.1.7)\\ \Gamma_0':\ \beta\frac{\partial u}{\partial \nu}+\gamma u=q & (14.1.8)\\ \Gamma_0:\ u=\bar{u} & (14.1.9)\end{cases}$$

即两者有相同的解。

这里变分问题的函数类 S 内的函数 u 当然默认有起码的光滑性，例如具有一阶导数 $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$ 以使积分(14.1.4)有意义，不去细说。关于等价性，仅列其论证要点如后，细节可参考[1]。

设对某函数 $u=u(x, y)$ 给以“变分”即增量 $\delta u=\delta u(x, y)$，函数从 u 变为 $u+\delta u$，则相应地 $J(u)$ 变为 $J(u+\delta u)$，不难用幂次展开的方法算出

$$J(u+\delta u)=J(u)+\delta J+\frac{1}{2}\delta^2 J \quad (14.1.10)$$

$$\delta J=\iint\limits_{\Omega}\left[\beta\frac{\partial u}{\partial x}\frac{\partial \delta u}{\partial x}+\beta\frac{\partial u}{\partial y}\frac{\partial \delta u}{\partial y}-f\delta u\right]dx\,dy+\int_{\Gamma_0'}[\gamma u-q]\,\delta u\,ds \quad (14.1.10')$$

$$\frac{1}{2}\delta^2 J=\iint\limits_{\Omega}\frac{1}{2}\left[\beta\left(\frac{\partial \delta u}{\partial x}\right)^2+\beta\left(\frac{\partial \delta u}{\partial y}\right)^2\right]dx\,dy+\int_{\Gamma_0'}\frac{1}{2}\eta(\delta u)^2\,ds \quad (14.1.10'')$$

这里视 δu 为无穷小量，δJ 线性地依赖于 u 又线性地依赖于 δu，因此为 δu 的同阶无穷小量，叫做泛函 J 的一次变分。$\delta^2 J$ 不依赖于 u 但二次地依赖于 δu，因此为 δu 的高阶无穷小量，叫做 J 的二次变分。

我们要求 u 及 $u+\delta u$ 都属于函数类 S，即都满足边界条件(14.1.2)，因此 δu 满足对应于(14.1.2)的齐次边界条件

$$\Gamma_0:\ \delta u=0 \tag{14.1.11}$$

满足这个边界条件的函数类记为 S_0 设 u 是 S 内某个特定函数，则 S 内的任意函数 v 必可表为 $v=u+\delta u$ 而 $\delta u\in S_0$。显然，对于一切 $\delta u\in S_0$ 恒有 $\delta^2 J\geqslant 0$。可以证明，在集合 Γ_0 非空（即确有第一类边界条件点）或者 $\eta\not\equiv 0$（即确有第三类边界条件点）的情况下，当 $\delta u\in S_0$ 而相应的 $\delta^2 J=0$ 时必有 $\delta u\equiv 0$，这就是说二次变分 $\delta^2 J$ 是正定的（见 14.1.2 节）。从 (14.1.10) 中各项的量级比较，可以证明，当二次变分 $\delta^2 J$ 为正定时（或半正定时）（椭圆型问题中绝大多数是这样的），函数 u 在 S 内使 J 达到极小的充要条件是一次变分 δJ 恒为零，即

$$\delta J=\delta J(u,\ \delta u)=0,\quad 对一切\ \delta u\in S_0 \tag{14.1.12}$$

现在来说明 (14.1.12) 与 (14.1.7～8) 等价。运用高斯积分公式

$$\begin{aligned}&\iint\limits_{\Omega}\left[\beta\frac{\partial u}{\partial x}\cdot\frac{\partial\delta u}{\partial x}+\beta\frac{\partial u}{\partial y}\cdot\frac{\partial\delta u}{\partial y}\right]dx\,dy\\&=-\iint\limits_{\Omega}\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}\right)\delta u\,dx\,dy+\int_{\partial\Omega}\beta\frac{\partial u}{\partial\nu}ds\end{aligned}$$

和边界条件 (14.1.11) 可得

$$\delta J=-\iint\limits_{\Omega}\left[\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+f\right]\delta u\,dx\,dy+\int_{\Gamma_1'}\left[\beta\frac{\partial u}{\partial\nu}+\eta u-q\right]\delta u\,ds=0$$

对一切 $\delta u\in S_0$。由 δu 的任意性可以推出上式两个积分号下的[…]恒为 0，这就是 (14.1.7～8)。反之当 (14.1.7～8) 成立时 (14.1.12) 必也成立，故 (14.1.12) 与 (14.1.7～8) 等价。因此变分问题 (14.1.5～6) 与边值问题 (14.1.7～9) 等价。

泛函的一次、二次变分实质上是普通多元函数的一次、二次变分的推广。事实上对于函数 $F(x_1,\ \cdots,\ x_n)$ 的变元 $x_1,\ \cdots,\ x_n$ 给以增量 $\delta x_1,\ \cdots,\ \delta x_n$ 则有

$$F(x_1+\delta x_1,\ \cdots,\ x_n+\delta x_n)=F(x_1,\ \cdots,\ x_n)+\delta F+\frac{1}{2}\delta^2 F \tag{14.1.13}$$

$$\delta F=\sum_{i=1}^{n}\frac{\partial F}{\partial x_i}\delta x_i$$

$$\frac{1}{2}\delta^2 F=\frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2 F}{\partial x_i\partial x_j}\delta x_i\delta x_j$$

δF, $\delta^2 F$ 便是 F 的一次、二次微分。在微积分中熟知有极值原理：当在某点 $(x_1,\ \cdots,\ x_n)$ 的二次微分 $\delta^2 F$——作为 $\delta x_1,\ \cdots,\ \delta x_n$ 的二次型为正定（相当于在 $(x_1,\ \cdots,\ x_n)$ 的二阶导数阵 $\dfrac{\partial^2 F}{\partial x_i\partial x_j}$ 为正定矩阵）时，该点 $(x_1,\ \cdots,\ x_n)$ 为 F 的极小点的充要条件为

$$\delta F=\delta F(x_1,\ \cdots,\ x_n;\ \delta x_1,\ \cdots,\ \delta x_n)=0,\quad 对一切\ \delta x_1,\ \cdots,\ \delta x_n \tag{14.1.14}$$

这也等价于

$$\frac{\delta F}{\partial x_1}=0,\ \cdots,\ \frac{\partial F}{\partial x_n}=0 \tag{14.1.15}$$

因此极值问题 $F(x_1,\ \cdots,\ x_n)=\min$ 等价于解方程组 (14.1.15) 的问题。以上 δJ, $\delta^2 J$ 分别对应于 δF, $\delta^2 F$, (14.1.12) 对应于 (14.1.14)、(14.1.7～8) 对应于 (14.1.15)，而变分问题 (14.1.5～6) 与边值问题 (14.1.7～9) 的等价性对应于函数 F 的极值问题与方程组 (14.1.15) 的等价性。

上面建立边值问题与变分问题的等价性时用了高斯积分公式，它仅当有关场量有一定的光滑性才是合法的，例如当系数β为连续函数时就是这样。当介质系数β有间断时，命其间断线为L(图 14.1 中的虚线)，它把区域Ω分割为几个子域，为简便计，设分为两块$\Omega=\Omega^-+\Omega^+$，在L上规定从Ω^-指向Ω^+的方向为法线上的方向，于是

$$\iint\limits_{\Omega}\cdots dx\,dy=\iint\limits_{\Omega^-}\cdots dx\,dy+\iint\limits_{\Omega^+}\cdots dx\,dy$$

在子域Ω^-, Ω^+内场量分别是光滑的，可以运用高斯积分公式，由此不难算出

$$\delta J=-\iint\limits_{\Omega-L}\left[\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+f\right]\delta u\,dx\,dy+\int\limits_{L}\left[\left(\beta\frac{\partial u}{\partial\nu}\right)^-\right.$$
$$\left.-\left(\beta\frac{\partial u}{\partial\nu}\right)^+\right]\delta u\,ds+\int_{\Gamma_0'}\left[\beta\frac{\partial u}{\partial\nu}+\gamma_1 u-q\right]\delta u\,ds=0$$

由此可以得出结论：在介质系数β有间断时，变分问题(14.1.5～6)等价于边值问题

$$\begin{cases}\Omega-L:\ \ -\left(\dfrac{\partial}{\partial x}\beta\dfrac{\partial u}{\partial x}+\dfrac{\partial}{\partial y}\beta\dfrac{\partial u}{\partial y}\right)=f\\ \quad L:\ \ \left(\beta\dfrac{\partial u}{\partial\nu}\right)^--\left(\beta\dfrac{\partial u}{\partial\nu}\right)^+=0\\ \quad \Gamma_0':\ \ \beta\dfrac{\partial u}{\partial\nu}+\gamma_1 u=q\\ \quad \Gamma_0:\ \ u=\bar{u}\end{cases}\tag{14.1.16}$$

这里与(14.1.7～9)比较，多出一个在间断线L上的交界条件

$$L:\ \left(\beta\frac{\partial u}{\partial\nu}\right)^-=\left(\beta\frac{\partial u}{\partial\nu}\right)^+\tag{14.1.17}$$

综上所述，微分方程(14.1.11)连同其第二、三类边界条件，以及介质系数有间断时的交界条件都可以从适当的变分原理导出。应该注意的是：在解微分方程时，第二、三类边界条件以及交界条件都必须作为定解条件列出；而在解相应的变分问题时，这些条件被极值函数自动满足，无须作为定解条件列出，因此称这类条件为自然边界条件。反之，第一类边界条件——如果有的话——在变分问题中与在微分方程问题中一样，必须作为定解条件列出，这类条件叫做强加边界条件。强加边界条件比较简单，在这里只涉及u本身，而自然边界条件则比较复杂，涉及到u以及法向导数$\dfrac{\partial u}{\partial\nu}$，当边界以及介质间断线的几何形状复杂时，处理是比较困难的。此外微分方程(14.1.11)中含有二阶导数，变分原理(14.1.5)中只含有一阶导数。因此，直接从变分原理出发来进行离散化和数值解是有利的。有限元法就是这样，由于它在离散化时采用了剖分插值方法，上述有利因素可以得到充分发挥。

变分问题与边值问题的等价性还可以推广到更复杂的情况，例如

$$\begin{cases}J(u)=\displaystyle\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\sum_{i,j=1}^{2}\beta_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}+\gamma u^2\right]-fu\right\}dx_1\,dx_2\\ \qquad\qquad+\displaystyle\int_{\Gamma_0'}\left[\frac{1}{2}\gamma_1 u^2-qu\right]ds=\text{极小}\\ \Gamma_0:\ u=\bar{u}\end{cases}\tag{14.1.18}$$

等价于

$$\begin{cases} \Omega - L: & -\left(\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} \beta_{ij} \frac{\partial u}{\partial x_j}\right) + \gamma u = f \\ L: & \left(\sum_{i,j=1}^{2} \beta_{ij} \nu_i \frac{\partial u}{\partial x_j}\right)^- - \left(\sum_{i,j=1}^{2} \beta_{ij} \nu_i \frac{\partial u}{\partial x_j}\right)^+ = 0 \\ \Gamma_0': & \sum_{i,j=1}^{2} \beta_{ij} \nu_i \frac{\partial u}{\partial x_j} + \eta u = q \\ \Gamma_0: & u = \bar{u} \end{cases} \tag{14.1.19}$$

这里将 x, y 记为 x_1, x_2, (ν_1, ν_2) 是法向余弦，$\beta_{ij} = \beta_{i,j}(x, y)$ 是对称正定阵，各向异性的介质系数就是作此形式，L 为系数 β_{ij} 的间断线。$\gamma \geqslant 0$，通常反映环境的反作用，例如在热传导问题中相当于介质与环境之间的热交换系数，在弹性力学中则相当于基础的弹性系数等等。可以见到，这里自然边界条件的形状更复杂，因此变分原理的有利因素更显著。

将(14.1.18~19)中的 Ω 理解为三维空间 x_1, x_2, x_3 中的立体，L, Γ_0, Γ_0' 了解为二维面，求和下标改为 i, $j=1$, 2, 3，则问题就推广到三维的情况。如将 Ω 理解为一维区间，Γ_0, Γ_0' 理解为边界点，L 为 Ω 内部的离散点，i, $j=1$，问题便简化为一维的，即

$$\begin{cases} J(u) = \int_\Omega \left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2 + \gamma u^2\right] - fu\right\} dx + \sum_{\Gamma_0'}\left[\frac{1}{2}\eta u^2 - qu\right] = \text{极小} \\ \Gamma_0: \ u = \bar{u} \end{cases} \tag{14.1.20}$$

等价于

$$\begin{cases} \Omega - L: & -\frac{\partial}{\partial x} \beta \frac{\partial u}{\partial x} + \gamma u = f \\ L: & \left(\beta \frac{\partial u}{\partial x}\right)^- - \left(\beta \frac{\partial u}{\partial x}\right)^+ = 0 \\ \Gamma_0': & \varepsilon \beta \frac{\partial u}{\partial x} + \eta u = q \\ \Gamma_0: & u = \bar{u} \end{cases} \tag{14.1.21}$$

这里 ε 在右边界上为 $+1$，在左边界上为 -1。

14.1.2 关于变分问题的正定性

对于变分问题，例如(14.1.5~6)，我们说二次变分 $\delta^2 J$ 是半正定的，如果

$$\delta^2 J(\delta u) \geqslant 0, \quad \text{对于一切 } \delta u \in S_0 \tag{14.1.22}$$

当系数 $\beta>0$, $\eta\geqslant 0$ 时，根据表达式(14.1.10″)显然可见(14.1.22)成立，即半正定。如果除了满足(14.1.22)以外还进一步满足

$$\delta u \in S_0, \ \delta^2 J(\delta u) = 0 \Rightarrow \delta u \equiv 0 \tag{14.1.23}$$

则称为正定的。如果这一补充条件不满足，也就是说存在 $\delta u \in S_0$, $\delta u \neq 0$ 而能使

$$\delta^2 J(\delta u) = 0$$

则称为退化半正定的。

现在来证明：当满足下列两条件之一时，$\delta^2 J$ 为正定。

(1) $\Gamma_0 \neq 0$(非空)，即边值问题确有第一类边界条件的区段。

(2) $\eta \neq 0$，即在 Γ_0' 上含有区段 $\Gamma_3 \neq 0$(非空)使得在 Γ_3 上 $\eta>0$，这就是说边值问题确有第三类边界条件的区段。

事实上，由于表达式(14.1.10)的积分号下都是正号的平方和，因此当 $\delta^2 J(\delta u) = 0$ 时

必有

$$\frac{\partial}{\partial x}\delta u\equiv\frac{\partial}{\partial y}\delta u=0\Rightarrow\delta u\equiv c=\text{常数},\text{在}\ \Omega\ \text{上。}$$

由于 $\delta u\in S_0$，故由(1)知在 Γ_0 上 $\delta u=0$，因此在 Ω 上 $\delta u\equiv c=0$。如果(1)不被满足，则由(2)知在 Γ_3 在 $\delta u=0$，因此同样有 $\delta u\equiv c=0$。

当条件(1)，(2)都不满足时实际上就是所谓第二类边值问题。这就是说 $\Gamma_0=0$ (空)即 $\partial\Omega=\Gamma_0'$ 并且 $\eta\equiv0$。这时变分问题(14.1.5～6)退化为无条件变分问题

$$J(u)=\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy-\oint_{\partial\Omega}qu\,ds=\text{极小}\qquad(14.1.24)$$

函数类 S 与 S_0 一致，边界上不受约束，而二次变分简化为

$$\frac{1}{2}\delta^2J(\delta u)=\iint\limits_{\Omega}\frac{1}{2}\left[\beta\left(\frac{\partial\delta u}{\partial x}\right)^2+\beta\left(\frac{\partial\delta u}{\partial y}\right)^2\right]dx\,dy\qquad(14.1.25)$$

等价的边值问题则成为第二类的，即

$$\begin{cases}\Omega-L:\ \ -\left(\dfrac{\partial}{\partial x}\beta\dfrac{\partial u}{\partial x}+\dfrac{\partial}{\partial y}\beta\dfrac{\partial u}{\partial y}\right)=f\\ L:\ \ \left(\beta\dfrac{\partial u}{\partial \nu}\right)^--\left(\beta\dfrac{\partial u}{\partial \nu}\right)^+=0\\ \partial\Omega:\ \ \beta\dfrac{\partial u}{\partial \nu}=q\end{cases}\qquad(14.1.26)$$

从上面在条件(1)，(2)下正定性的论证中可以看到，当条件(1)，(2)都不成立，即对于(14.1.24)的情况，二次变分是退化的，即

$$\delta^2J(\delta u)=0\Leftrightarrow\delta u\equiv c=\text{常数}$$

在正定即非退化的情况，问题(14.1.5～6)即(14.1.16)有唯一解，而在退化的情况，问题(14.1.24)即(14.1.26)可以没有解，有解时也不唯一。这是一个很大的区别。在退化的情况，可以证明(参看第十三章§13.1)：

1. 齐次问题，即 $f\equiv0$，$q\equiv0$ 的情况，也就是

$$J(u)=\iint\limits_{\Omega}\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy=\text{极小}\qquad(14.1.27)$$

或

$$\begin{cases}\Omega-L:\ \ -\left(\dfrac{\partial}{\partial x}\beta\dfrac{\partial u}{\partial x}+\dfrac{\partial}{\partial y}\beta\dfrac{\partial u}{\partial y}\right)=0\\ L:\ \ \left(\beta\dfrac{\partial u}{\partial \nu}\right)^--\left(\beta\dfrac{\partial u}{\partial \nu}\right)^+=0\\ \partial\Omega:\ \ \beta\dfrac{\partial u}{\partial \nu}=0\end{cases}\qquad(14.1.28)$$

有无穷多非零解，可以表为

$$u\equiv c=\text{常数}\qquad(14.1.29)$$

2. 非齐次问题(14.1.24)或(14.1.26)有解的充要条件是

$$\iint\limits_{\Omega}f\,dx\,dy+\oint_{\partial\Omega}q\,ds=0\qquad(14.1.30)$$

通称为协调条件。当协调条件(14.1.29)被满足时，非齐次问题(14.1.24)或(14.1.26)的通解 u 可以表为一个特解 $\tilde{u}$ 和相齐齐次问题(14.1.27)或(14.1.28)的通解之和，即

$$u=\tilde{u}+c \tag{14.1.31}$$

协调条件(14.1.30)在物理上是自然的。例如，把(14.1.26)理解为不受约束的薄膜平衡问题，f, q 为外载荷，条件(14.1.30)表示外载荷达成平衡，显然只有当外载荷本身达成平衡时，不受约束的薄膜才可能达成平衡。齐次问题的非零解(14.1.29)相当于刚性位移为不受约束不受载荷的薄膜的平衡位移。

上面所举的变分问题中，二次泛函都是半正定，包括正定或退化，能量 J 所达到的极值是极小值。在有限元方法中大多数实际问题都属此类。但是也有一些实际问题中人们要求的并不是能量 J 达到极小值而只是所谓临界值，这时二次泛函为不定即 $\delta^2 J$ 可以有正值也可以有负值。在微积分中，对于多元函数 $F(x_1, \cdots, x_n)$ 满足 $\frac{\partial F}{\partial x_1}=0, \cdots, \frac{\partial F}{\partial x_n}=0$ 的点 $(x_1, \cdots, x_n)$ 的点叫做临界点(或逗留点)，相应的 F 值叫做临界值(或逗留值)，而不论二阶导数阵 $\frac{\partial^2 F}{\partial x_i \partial x_j}$ 是否正定，故临界值可以是极小值，也可以是极大值，也可能都不是。类似地，我们说 $J(u)$ 在函数类 S 中达到临界(或逗留)是指其一次变分恒为零，即(14.1.12)成立而不论二次变分 $\delta^2 J$ 是否正定或半正定。在我们所讨论的典型例中，变分问题与边值问题的等价性主要是通过临界性建立起来的。不过由于二次变分的正定性，临界性逐成为极小性。一般说来，变分问题可以根据其二次变分为正定或不定(即 $\delta^2 J$ 对于某些 δu 取正值，对于另一些 δu 取负值)而分为两类。椭圆方程中的势能原理多属于正定型的。椭圆方程中的余能原理以及双曲方程中的最小作用原理等则多属于不定型。正定性对于变分原理的误差估计和收敛性论证是关键的；对于离散化后代数问题的解算也是有利的。但在不定的情况下也并不妨碍变分原理的实际运用，这是因为变分原理这套形式工具在计算实践中所起的作用，如果边界条件的自动实现以及导数的降阶等方面主要是由临界性带来的，与正定性无关。

§14.2 几何剖分与分片插值

14.2.1 三角剖分

对于平面区域作剖分时，基本单元可取为三角形、矩形、四边形、曲边的多边形等等或兼而有之。单纯的三角形剖分最简单常用，适应性较强，因此只介绍这一种。

设有平面域 Ω，如果 Ω 的边界 $\partial\Omega$ 是曲的，则总可以裁弯取直，用适当的折线来逼近，这样 Ω 就近似地代以一个多边形域，仍记作 Ω。

把多边形域 Ω 剖分为一系列的三角形，更确切地说，剖分为

点元：$A_1, A_2, \cdots, A_{N_0}$

线元：$B_1, B_2, \cdots, B_{N_1}$

面元：$C_1, C_2, \cdots, C_{N_2}$

N_0, N_1, N_2 为点、线、面元的个数。面元是三角形，线元是直线段。每个面元以三个线元为其边，也以三个点元为其顶点，每个线元以两个点元为其端点即顶点。如果区域 Ω 的内部和边界上的介质系数如 β, η 以及 f, q 有间断性，则间断的线、点应落在线元和点元上，也就是说剖分应与问题本有的分割相协调，图 14.1 中虚线表示内部的间断线，图 14.2 表示对应于图 14.1 的一个剖分，其中对应于间断线的线元仍用虚线表示。

如上把点元、线元、面元都上了号并给出

1．点元坐标(x_k, y_k)，$k=1, 2, \cdots, N_0$;

2．线元两顶点的编号(m_{1k}, m_{2k})，$k=1, 2, \cdots, N_1$;

3．面元三顶点的编号(n_{1k}, n_{2k}, n_{3k})，$k=1, 2, \cdots, N_2$;

则剖分就完全确定。

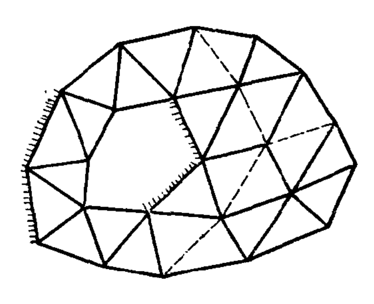


图 14.2

对于区域的剖分，除了如上所述必须与问题的物理条件的划分相协调外，基本上可以是任意的，可以在关键或关心的部位加密，在另外的部位放疏。这种灵活性是有限元法的一个优点。但是也要注意：(1)不要有太"扁"的三角形，即避免出现最小内角接近于$0°$的三角形；(2) 剖分疏密的过渡不要太陡。不然的话，会引起离散后代数方程组系数矩阵的病态，不利于解算，并且影响到精确度和收敛性。

对平面域Ω作三角剖分时，点、线、面元的个数N_0, N_1, N_2一般是任意的，但它们之间有一定的比例关系。首先，有尤拉公式

$$N_0-N_1+N_2=1-p$$

p为域Ω的孔数，单连通时$p=0$；这一公式不限于三角剖分，对其它剖分也成立。它表示，不论怎样剖分，$N_0-N_1+N_2$恒不变，是区域Ω的一个拓扑不变量。此外，每个三角元以三个线元为边，每个线元邻接两个(当它在内部时)或一个(当它在边界上时)面元，因此

$$3N_2=2N_1-N_1'$$

N_1'为边界线元的个数。在计算实践上，当剖分较细时，恒有$N_1\gg N_1'$, N_0, N_1, $N_2\gg 1-p$，因此$3N_2\approx 2N_1$, $N_1\approx\frac{3}{2}N_2$, $N_0-N_1+N_1\approx 0$, $N_0\approx N_1-N_2\approx\frac{3}{2}N_2-N_2\approx\frac{1}{2}N_2$，因此有近似的比例关系

$$N_0:N_2:N_1\approx 1:2:3$$

它仅对三角剖分成立。

14.2.2 三角形上的线性插值

在有限元的离散化中，特解函数$u(x, y)$在各个单元上用适当的插值函数来代替，最简单的插值方法就是三角形上的线性插值，不仅它被广泛应用，同时也是其它三角形上插值方法的基础。

设有任意三角形$C=(A_1, A_2, A_3)$，顶点A_i的坐标为(x_i, y_i)，$i=1, 2, 3$，设有某函数$u(x, y)$在顶点的值为$u_i=u(x_i, y_i)$，$i=1, 2, 3$，要求作线次函数即一次多项式

$$U(x, y)=ax+by+c$$

使得

$$U(x_1, y_1)=ax_1+by_1+c=u_1$$
$$U(x_2, y_2)=ax_2+by_2+c=u_2$$
$$U(x_3, y_3)=ax_3+by_3+c=u_3$$

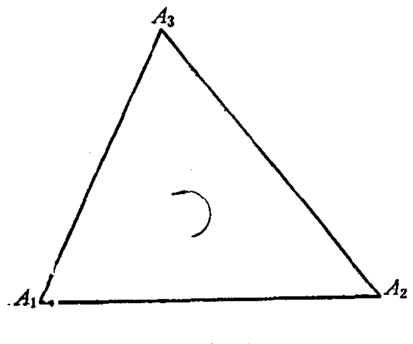


图 14.3

为了方便，在这里以及以后各节恒采用下列统一的记号

$$\xi_1=x_2-x_3,\qquad \xi_2=x_3-x_1,\qquad \xi_3=x_1-x_2$$

$$\eta_1=y_2-y_3,\qquad \eta_2=y_3-y_1,\qquad \eta_3=y_1-y_2$$

$$\omega_1=x_2y_3-x_3y_2,\ \omega_2=x_3y_1-x_1y_3,\ \omega_3=x_1y_2-x_2y_1$$

$$D=\begin{vmatrix} x_1 & y_1 & 1\\ x_2 & y_2 & 1\\ x_3 & y_3 & 1\end{vmatrix}=\xi_1\eta_2-\xi_2\eta_1=\xi_2\eta_3-\xi_3\eta_2=\xi_3\eta_1-\xi_1\eta_3=\omega_1+\omega_2+\omega_3 \tag{14.2.1}$$

$$D_0=|D|,\quad \varepsilon=\operatorname{sign}D=\begin{cases}1, & 当\ D>0\\ +1, & 当\ D<0\end{cases}$$

顺便指出，当 $A_1,\ A_2,\ A_3$ 作逆时针向(如图 14.3)时 $D>0$，作顺时针向时 $D<0$。三角形 C 的面积(恒正)为

$$\iint_C dx\,dy=D_0/2$$

由前列的三个插值方程可以解出

$$a=\frac{1}{D}\begin{vmatrix} u_1 & y_1 & 1\\ u_2 & y_2 & 1\\ u_3 & y_3 & 1\end{vmatrix}=\frac{1}{D}\sum_{i=1}^{3}\eta_i u_i$$

$$b=\frac{1}{D}\begin{vmatrix} x_1 & u_1 & 1\\ x_2 & u_2 & 1\\ x_3 & u_3 & 1\end{vmatrix}=-\frac{1}{D}\sum_{i=1}^{3}\xi_i u_i$$

$$c=\frac{1}{D}\begin{vmatrix} x_1 & y_1 & u_1\\ x_2 & y_2 & u_2\\ x_3 & y_3 & u_3\end{vmatrix}=\frac{1}{D}\sum_{i=1}^{3}\omega_i u_i$$

于是

$$U(x,\ y)=\frac{1}{D}\left(x\sum_{i=1}^{3}\eta_i u_i-y\sum_{i=1}^{3}\xi_i u_i+\sum_{i=1}^{3}\omega_i u_i\right)$$

为了方便，命

$$\lambda_i(x,\ y)=(\eta_i x-\xi_i y+\omega_i)/D,\quad i=1,\ 2,\ 3 \tag{14.2.2}$$

于是，以节点值 $u_1,\ u_2,\ u_3$ 为基础的线性插值函数 U 可以表为

$$U(x,\ y)=\sum_{i=1}^{3}u_i\lambda_i(x,\ y) \tag{14.2.3}$$

函数 $\lambda_i=\lambda_i(x,\ y)$ 可以称为三角形上线性插值的基函数，它们本身也是线性函数并满足

$$\lambda_i(x_j,\ y_j)=\delta_{ij}=\begin{cases}1, & 当\ i=j\\ 0, & 当\ i\neq j\end{cases}$$

顺便指出，当被插出数 $u(x,\ y)$ 自己是线性函数时，它的三顶点线性插值是准确的，即 $u(x,\ y)=U(x,\ y)$。因此，依次取 $u(x,\ y)=1,\ x,\ y$ 时即得恒等式

$$1\equiv\lambda_1+\lambda_2+\lambda_3 \tag{14.2.4}$$

$$x\equiv x_1\lambda_1+x_2\lambda_2+x_3\lambda_3 \tag{14.2.5}$$

$$y=y_1\lambda_1+y_2\lambda_2+y_3\lambda_3 \tag{14.2.6}$$

基函数 λ_i 是线性的，它们的偏导数是常数

$$\begin{cases} \dfrac{\partial}{\partial x}\lambda_i(x,\ y)=\eta_i/D \\ \dfrac{\partial}{\partial y}\lambda_i(x,\ y)=-\xi_i/D \end{cases} \tag{14.2.7}$$

$$\frac{\partial U}{\partial x}=\sum_{i=1}^{3}u_i\frac{\partial\lambda_i}{\partial x}=\sum_{i=1}^{3}u_i\eta_i/D$$

$$\frac{\partial U}{\partial y}=\sum_{i=1}^{3}u_i\frac{\partial\lambda_i}{\partial y}=-\sum_{i=1}^{3}u_i\xi_i/D$$

在有限元法计算能量表达式时要用到 λ_i 及其导数的乘积的积分。根据积分公式(见14.2.4节)

$$\iint_C \lambda_1^{p_1}\lambda_2^{p_2}\lambda_3^{p_3}dx\,dy=\frac{p_1!\,p_2!\,p_3!}{(p_1+p_2+p_3+2)!} \tag{14.2.8}$$

以及(14.2.7)可以得出下列积分表

表 14.1 $\iint_C \varphi\psi\,dx\,dy$

φ \ ψ	1	λ_j	$\dfrac{\partial\lambda_j}{\partial x}$	$\dfrac{\partial\lambda_j}{\partial y}$
1	$D_0/2$			
λ_i	$D_0/6$	$D_0(1+\delta_{ij})/24$		
$\dfrac{\partial\lambda_i}{\partial x}$	$e\eta_i/2$	$e\eta_i/6$	$\eta_i\eta_j/2D_0$	
$\dfrac{\partial\lambda_i}{\partial y}$	$-e\xi_i/2$	$-e\xi_i/6$	$-\xi_i\eta_j/2D_0$	$\xi_i\xi_j/2D_0$

有时,例如对轴对称问题,需要用到如 $\iint\cdots x\,dx\,dy$ 形状的积分,命

$$x_0=\frac{1}{3}(x_1+x_2+x_3)$$

于是根据(14.2.8)和同上的方法可以得下表

表 14.2 $\iint_C \varphi\psi x\,dx\,dy$

φ \ ψ	1	λ_j	$\dfrac{\partial\lambda_j}{\partial x}$	$\dfrac{\partial\lambda_j}{\partial y}$
1	$x_0D_0/2$			
λ_i	$(3x_0+x_i)D_0/24$	$(3x_0+x_i+x_j)D_0(1+\delta_{ij})/120$		
$\dfrac{\partial\lambda_i}{\partial x}$	$e\eta_ix_0/2$	$e\eta_i(3x_0+x_j)/24$	$\eta_i\eta_jx_0/2D_0$	
$\dfrac{\partial\lambda_i}{\partial y}$	$-e\xi_ix_0/2$	$-e\xi_i(3x_0+x_j)/24$	$-\xi_i\eta_jx_0/2D_0$	$\xi_i\xi_jx_0/2D_0$

14.2.3 线元上的线性插值

取三角形 $C=(A_1, A_2, A_3)$的一个任意边，例如线元$(A_1, A_2)=B$，命 S 为自 A_1 至 A_2 的弦长变量，C 上的函数 $u(x, y)$，及其插值 $U(x, y)$ 以及基函数 $\lambda_i(x, y)$ 在 B 的值记为 $u(s)$, $U(s)$, $\lambda_i(s)$；由于 λ_3 在 A_3 的对边即(A_1, A_2)上恒为 0，故有

$$U(s)=\sum_{i=1}^{2} u_i\lambda_i(s) \tag{14.2.9}$$

因此，在 B 上，$U(s)$就是由 $u(s)$在两端的值 u_1, u_2 所产生的线性插值，与第三个顶点值 u_3 无关。因此在线元 B 上,独立地用两顶点的线性插值与以 B 为一边的面元 C 的三顶点线性插值的结果是一致的。

命 L 为线元 B 的长度

$$L=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{14.2.10}$$

显然有

$$\lambda_1(s)=1-\frac{s}{L}, \quad \lambda_2(s)=\frac{s}{L} \tag{14.2.11}$$

$$\frac{\partial\lambda_1}{\partial s}=-1/L, \quad \frac{\partial\lambda_2}{\partial s}=1/L, \quad 即 \quad \frac{\partial\lambda_i}{\partial s}=(-1)^i/L \tag{14.2.12}$$

与三角面元相类似,在线元 $B=(A_1, A_2)$上有下列公式

$$\lambda_i(x_j, y_j)=\delta_{ij}, \quad i, j=1, 2$$

$$1\equiv\lambda_1+\lambda_2$$

$$x\equiv x_1\lambda_1+x_2\lambda_2$$

$$y\equiv y_1\lambda_1+y_2\lambda_2$$

$$\int_B \lambda_1^{p_1}\lambda_2^{p_2}ds=\frac{p_1!p_2!L}{(p_1+p_2+1)!} \tag{14.2.13}$$

由此,并命

$$x_0=\frac{1}{2}(x_1+x_2)$$

则得下列积分表。

表 14.3 $\int_B \varphi\psi\, ds$

φ \ ψ	1	λ_j	$\frac{\partial\lambda_j}{\partial s}$
1	L		
λ_i	$L/2$	$L(1+\delta_{ij})/6$	
$\frac{\partial\lambda_i}{\partial s}$	$(-1)^i$	$(-1)^i/2$	$(-1)^{i+j}/L$

表 14.4 $\int_B \varphi\psi x\, ds$

φ \ ψ	1	λ_j	$\frac{\partial\lambda_j}{\partial s}$
1	x_0L		
λ_i	$(2x_0+x_i)L/6$	$(x_0+x_i\delta_{ij})L/6$	
$\frac{\partial\lambda_i}{\partial s}$	$(-1)^ix_0$	$(-1)^i(2x_0+x_j)/6$	$(-1)^{i+j}x_0/L$

总结上述，对于 Ω 上的函数 $u(x, y)$，按照三角剖分分别在每个面元作线性插值，它们在相邻面元的公共边及公共点上取相同的值，因此拼起来得到在 Ω 上的分片线性插值函数 $U(x, y)$，系由 $u(x, y)$在各顶点 $A_1, \cdots, A_{N_0}$ 处的值 $u_1, \cdots, u_{N_0}$ 决定，$U(x, y)$在每个单元(面、线、点)上就是有关顶点 u 值的线性插值。$U(x, y)$在 Ω 上整体上是连续函数；但一阶

导数有间断,实际上是分片常数。

14.2.4 重心坐标

在三角形 $C=(A_1, A_2, A_3)$ 上作插值和微积分运算时，线性函数 $\lambda_1, \lambda_2, \lambda_3$ 占有重要地位。给了一个点 P 的直角坐标 (x, y)，用公式(14.2.2)可以算出相应的 $(\lambda_1, \lambda_2, \lambda_3)$，$\lambda_1, \lambda_2, \lambda_3$ 中只有两个是独立的,它们满足恒等式(14.2.4)

$$\lambda_1+\lambda_2+\lambda_3=1$$

反之给了满足这一等式的三个数 $(\lambda_1, \lambda_2, \lambda_3)$ 则用(14.2.5~6)可以算出相应的 (x, y)。因此 $(\lambda_1, \lambda_2, \lambda_3)$ 和 (x, y) 一样可以作为坐标，通常叫做重心坐标。这是因为，取三个质量 $\lambda_1, \lambda_2, \lambda_3$，其和为1，分别放在顶点 A_1, A_2, A_3 上,则这个质量系统的重心 $P(x, y)$ 就是

$$x=\frac{\lambda_1x_1+\lambda_2x_2+\lambda_3x_3}{\lambda_1+\lambda_2+\lambda_3}=\lambda_1x_1+\lambda_2x_2+\lambda_3x_3$$

$$y=\frac{\lambda_1y_1+\lambda_2y_2+\lambda_3y_3}{\lambda_1+\lambda_2+\lambda_3}=\lambda_1y_1+\lambda_2y_2+\lambda_3y_3$$

这就是公式(14.2.5~6)。

$(\lambda_1, \lambda_2, \lambda_3)$ 也叫做面积坐标。事实上，设点 $P=(x, y)$ 位于三角形 $A_1A_2A_3$ 之内并设 $A_1A_2A_3$ 作逆时针向，于是 PA_2A_3, PA_3A_1, PA_1A_2 也都作逆时针向(见图 14.4)，并有面积公式

$$S_0=\triangle A_1A_2A_3=\frac{1}{2}\begin{vmatrix}x_1 & y_1 & 1\\ x_2 & y_2 & 1\\ x_3 & y_3 & 1\end{vmatrix}=\frac{1}{2}D$$

$$S_1=\triangle PA_2A_3=\frac{1}{2}\begin{vmatrix}x & y & 1\\ x_2 & y_2 & 1\\ x_3 & y_3 & 1\end{vmatrix}=\frac{1}{2}(\eta_1x-\xi_1y+\omega_1)=\lambda_1S_0$$

$$S_2=\triangle PA_3A_1=\frac{1}{2}\begin{vmatrix}x & y & 1\\ x_3 & y_3 & 1\\ x_1 & y_1 & 1\end{vmatrix}=\frac{1}{2}(\eta_2x-\xi_2y+\omega_2)=\lambda_2S_0$$

$$S_3=\triangle PA_1A_2=\frac{1}{2}\begin{vmatrix}x & y & 1\\ x_1 & y_1 & 1\\ x_2 & y_2 & 1\end{vmatrix}=\frac{1}{2}(\eta_3x-\xi_3y+\omega_3)=\lambda_3S_0$$

因此对应于 $P=(x, y)$ 的 λ_i 就是面积比 S_i/S_0, $i=1, 2, 3$

$$\lambda_1=\triangle PA_2A_3/\triangle A_1A_2A_3,\ \lambda_2=\triangle PA_3A_1/\triangle A_1A_2A_3,\ \lambda_3=\triangle PA_1A_2/\triangle A_1A_2A_3$$

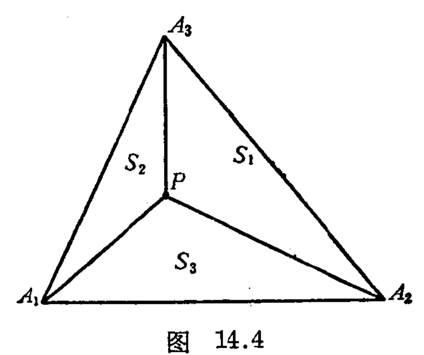


图 14.4

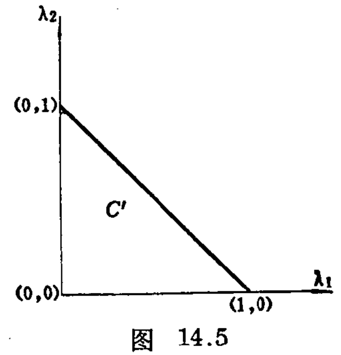


图 14.5

当点 $P=(x, y)$ 在三角形 C 上变时，相应的 $(\lambda_1, \lambda_2, \lambda_3)$ 的变化范围是

$$0\leqslant\lambda_1,\ \lambda_2,\ \lambda_3\leqslant 1, \quad \lambda_1+\lambda_2+\lambda_3=1$$

在 A_i 的对边上 $\lambda_i=0(i=1, 2, 3)$，A_1, A_2, A_3 的重心坐标则是 $(1, 0, 0)$; $(0, 1, 0)$; $(0, 0, 1)$。

如取 λ_1, λ_2 为独立变量，$\lambda_3=1-\lambda_1-\lambda_2$，2·∫(14.2.2) 把 xy 平面上的三角形 $C=(A_1, A_2, A_3)$ 变为 $\lambda_1\lambda_2$ 平面上的三角形 C': $0\leqslant\lambda_1, \lambda_2\leqslant 1$, $\lambda_1+\lambda_2\leqslant 1$ 如图 14.5 这个变换的导数行列式就是

$$\frac{\partial(\lambda_1, \lambda_2)}{\partial(x, y)}=\begin{vmatrix}\dfrac{\partial\lambda_1}{\partial x} & \dfrac{\partial\lambda_1}{\partial y}\\ \dfrac{\partial\lambda_2}{\partial x} & \dfrac{\partial\lambda_2}{\partial y}\end{vmatrix}=\frac{1}{D^2}\begin{vmatrix}\eta_1 & -\xi_1\\ \eta_2 & -\xi_2\end{vmatrix}=\frac{1}{D} \quad \frac{\partial(x, y)}{\partial(\lambda_1, \lambda_2)}=D$$

在三角形上求积分时，变到重心坐标较便，特别当被积函数本身 C 用重心坐标表示时：

$$\iint\limits_C F(\lambda_1(x, y), \lambda_2(x, y), \lambda_3(x,y))\,dx\,dy=\iint\limits_{C'} F(\lambda_1, \lambda_2, 1-\lambda_1-\lambda_2)\left|\frac{\partial(x, y)}{\partial(\lambda_1, \lambda_2)}\right| d\lambda_1\,d\lambda_2$$

$$=D_0\int_0^1 d\lambda_2\int_0^{1-\lambda_2} F(\lambda_1, \lambda_2, 1-\lambda_1-\lambda_2)\,d\lambda_1$$

表 14.5 三角形上数值积分公式

$$\int_C F(\lambda_1, \lambda_2, \lambda_3)\,ds=\frac{D_0}{2}\sum_{k=1}^{m}\rho^{(k)}F(\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_3^{(k)}),\quad \frac{D_0}{2}=C \text{ 的面积}$$

节点个数 m	节点坐标 $(\lambda_1, \lambda_2, \lambda_3)$	权数 ρ	精度次数 n
1	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	1	1
3	$\left(0, \frac{1}{2}, \frac{1}{2}\right)$, $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$, $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$	$\frac{1}{3}$	1
7	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	$\frac{27}{60}$	3
	$\left(0, \frac{1}{2}, \frac{1}{2}\right)$, $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$, $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$	$\frac{8}{60}$	
7	$(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$	$\frac{3}{60}$	5
	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	0.225	
$\alpha_1=0.05961587$ $\beta_1=0.47014206$	$(\alpha_1, \beta_1, \beta_1)$, $(\beta_1, \alpha_1, \beta_1)$, $(\beta_1, \beta_1, \alpha_1)$	0.13239415	
$\alpha_2=0.79742699$ $\beta_2=0.10128651$	$(\alpha_2, \beta_2, \beta_2)$, $(\beta_2, \alpha_2, \beta_2)$, $(\beta_2, \beta_2, \alpha_2)$	0.12593918	

据此，并利用尤拉积分

$$\int_0^1 s^n(1-s)^m ds=\frac{m!n!}{(m+n+1)!}$$

就可以导出公式(14.2.8)。

在有限元法中有时要用三角形上的数值积分，这也可以用重心坐标来表示，其一般形式为

$$\iint_C F(\lambda_1,\ \lambda_2,\ \lambda_3)dx\,dy=\frac{D_0}{2}\sum_{k=1}^{m}\rho^{(k)}F(\lambda_1^{(k)},\ \lambda_2^{(k)},\ \lambda_3^{(k)})$$

$D_0/2$ 是三角形 C 的面积，$(\lambda_1^{(k)},\ \lambda_2^{(k)},\ \lambda_3^{(k)})$是一组特定的节点，$\rho^{(k)}$ 是相应的权数。列举几种常用的公式如表 14.5，其中精度次数 n 是指公式对于 n 次多项式为准确的。

在线元 $B=(A_1, A_2)$上有一个弦长坐标 s(自 A_1 指向 A_2)，两个重心坐标(λ_1, λ_2)，$\lambda_1+\lambda_2=1$(见 14.2.3 节)。当点 $P=(s)$在 B 上变时，相应的$(\lambda_1,\ \lambda_2)$的变化范围是

$$0\leqslant\lambda_1,\ \lambda_2\leqslant 1,\quad \lambda_1+\lambda_2=1$$

A_1, A_2 的重心坐标是(1, 0), (0, 1)。取 λ_1 为独立变量 $\lambda_2=1-\lambda_1$，则当 $P=(s)$在 B 上变时 λ_1 在区间[0, 1]上变；显然有积分公式

$$\int_B F(\lambda_1(s),\lambda_2(s))ds=\int_0^1 F(\lambda_1,\ 1-\lambda_1)\left|\frac{ds}{d\lambda_1}\right|d\lambda_1=L\int_0^1 F(\lambda_1,\ 1-\lambda_1)d\lambda_1$$

L 为 B 的长度。据此以及尤拉积分就得积分公式(14.2.13)。

线元上的数值积分公式也可以用重心坐标来表示，其一般形式为

$$\int_B F(\lambda_1,\ \lambda_2)ds=L\sum_{k=1}^{m}\rho^{(k)}F(\lambda_1^{(k)},\ \lambda_2^{(k)})$$

$(\lambda_1^{(k)},\ \lambda_2^{(k)})$为线元上一组特定的节点，$\rho^{(k)}$ 为相应的权数。列举几种公式如表 14.6，实质上就是普通的辛浦生公式和高斯型公式。

表 14.6　线元上的数值积分公式

$\int_B F(\lambda_1,\ \lambda_2)ds=\sum^m\rho^{(k)}F(\lambda_1^{(k)},\ \lambda_2^{(k)})$，$L=B$ 的长度			
节点个数 m	节点坐标 $(\lambda_1,\ \lambda_2)$	权数 ρ	精度次数 n
1	$\left(\frac{1}{2},\ \frac{1}{2}\right)$	1	1
2 $\alpha_1=0.2113248654$ $1-\alpha_1=0.7886751346$	$(\alpha_1,\ 1-\alpha_1)$ $(1-\alpha_1,\ \alpha_1)$	$\frac{1}{2}$	3
3	$\left(\frac{1}{2},\ \frac{1}{2}\right)$ (0, 1) (1, 0)	$\frac{4}{6}$ $\frac{1}{6}$	3
3 $\alpha_2=0.1127016654$ $1-\alpha_2=0.8872983346$	$\left(\frac{1}{2},\ \frac{1}{2}\right)$ $(\alpha_2,\ 1-\alpha_2)$ $(1-\alpha_2,\ \alpha_2)$	$\frac{8}{18}$ $\frac{5}{18}$	5

14.2.5 三角形上的二次插值

三角形上的三点线性插值是最简单的插值方法，但只有起码的精度。可以构造较高精度的插值，首先就是六点二次插值。

在三角面元 $C=(A_1, A_2, A_3)$ 上取六个点作为插值节点，即三个顶点 A_1, A_2, A_3，简记为1, 2, 3, 以及它们的对边的中点，简记为4, 5, 6, 恒约定1与4, 2与5, 3与6相对，见图14.6。

函数 $u(x, y)$ 在各节点 (x_i, y_i) 的值记为 u_i, $i=1, \cdots, 6$, 要求定出完整的二次多项式

$$U(x, y)=a_1+a_2x+a_3y+a_4x^2+a_5xy+a_6y^2$$

满足

$$U(x_i, y_i)=u_i, \quad i=1, \cdots, 6$$

可以解出六个系数 $a_1, \cdots, a_6$。演算从略，只给出最终的结果如下

$$U(x, y)=\sum_{i=1}^{6} u_i\varphi_i(x, y) \qquad (14.2.14)$$

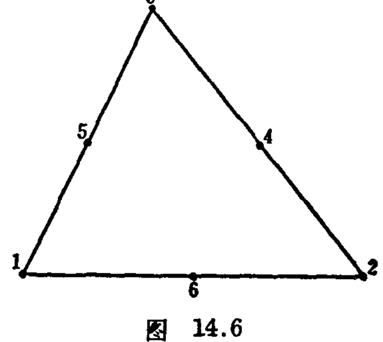


图 14.6

这里 φ_i 可以通过重心坐标 $\lambda_1, \lambda_2, \lambda_3$ 表达

$$\begin{cases} \varphi_i(x, y)=2\lambda_i^2-\lambda_i, \quad i=1, 2, 3 \\ \varphi_4(x, y)=4\lambda_2\lambda_3, \ \varphi_5(x, y)=4\lambda_3\lambda_1, \ \varphi_6(x, y)=4\lambda_1\lambda_2 \end{cases} \qquad (14.2.15)$$

φ_i 都是二次多项式。如果注意到在1, 2, 3, 4, 5, 6各节点的重心坐标 $(\lambda_1, \lambda_2, \lambda_3)$ 为 $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, $\left(0, \frac{1}{2}, \frac{1}{2}\right)$, $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$, $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$, 则容易验证

$$\varphi_i(x_j, y_j)=\delta_{ij}, \quad i, j=1, \cdots, 6$$

因此公式(14.2.15)确实就是所要求的二次插值函数。

根据复合微商法则以及 λ_i 的微商公式(14.2.7)可得 U 及 φ_i 的一阶微商，它们都是线性的。

$$\begin{cases} \dfrac{\partial}{\partial x}U=\sum_{i=1}^{6}u_i\dfrac{\partial\varphi_i}{\partial x}, \quad \dfrac{\partial}{\partial y}U=\sum_{i=1}^{6}u_i\dfrac{\partial\varphi_i}{\partial y} \\ \dfrac{\partial\varphi_i}{\partial x}=\eta_i(4\lambda_i-1), \quad i=1, 2, 3 \\ \dfrac{\partial\varphi_4}{\partial x}=4(\eta_3\lambda_2+\eta_2\lambda_3), \ \dfrac{\partial\varphi_5}{\partial x}=4(\eta_1\lambda_3+\eta_3\lambda_1), \ \dfrac{\partial\varphi_6}{\partial x}=4(\eta_2\lambda_1+\eta_1\lambda_2) \\ \dfrac{\partial\varphi_i}{\partial y}=-\xi_i(4\lambda_i-1), \quad i=1, 2, 3 \\ \dfrac{\partial\varphi_4}{\partial y}=-4(\xi_3\lambda_2+\xi_2\lambda_3), \ \dfrac{\partial\varphi_5}{\partial y}=-4(\xi_1\lambda_3+\xi_3\lambda_1), \ \dfrac{\partial\varphi_6}{\partial y}=-4(\xi_2\lambda_1+\xi_1\lambda_2) \end{cases} \qquad (14.2.16)$$

在三角形 C 的任意边，例如线元 $B=(A_1, A_2)$ 上，由于在此边上 $\lambda_3\equiv0$, 因此

$$U(s)=u_1\varphi_1(s)+u_2\varphi_2(s)+u_6\varphi_6(s)$$

点6就是 A_1A_2 的中点，s 表示自 A_1 至 A_2 的弦长参数(14.2.3节)。因此，面元上的二次插值在每个边上的值就是通过该边的三个节点(两个端点，一个中点)的二次插值而不依赖于位于此边以外的其它节点的值。这样，按三角面元分别作二次插值后，拼起来得到在 Ω 上

分片二次，整体为连续的函数，一阶导数为分片一次，但一般有间断。

将 $B=(A_1, A_2)$ 的两端点记为 1, 2, 并将其中点改记为 3, 于是

$$U(s)=\sum_{i=1}^{3}u_i\varphi_i(x) \tag{14.2.17}$$

$$\begin{cases}\varphi_i(s)=2\lambda_i^2-\lambda_i, & i=1,\ 2\\ \varphi_3(s)=4\lambda_1\lambda_2\end{cases} \tag{14.2.18}$$

这里 $\lambda_i=\lambda_i(s)$，见 14.2.3 节。相应的导数公式为

$$\begin{aligned}&\frac{\partial U}{\partial s}=\sum_{i=1}^{3}u_i\frac{\partial\varphi_i}{\partial s}\\&\frac{\partial\varphi_i}{\partial s}=(-1)^i(4\lambda_i-1)/L, \quad i=1,\ 2\\&\frac{\partial\varphi_3}{\partial s}=4(\lambda_1-\lambda_2)/L\end{aligned} \tag{14.2.19}$$

将分片二次插值用于有限元法时，需要插值基函数 φ_i 及其导数的乘积的积分。可以根据它们的表达式以及积分公式(14.2.8)，(14.2.13)算出类似于表 14.1~4 那样的表格，但比较庞杂故不列。事实上，在程序实现时，宁可采用如 14.2.4 节中所介绍的数值积分方法，反而简便，并且有更大的通用性。

§ 14.3 变分问题的离散化

我们将以变分问题(14.1.5~6)为例说明能量积分的离散化。为了说明一般的原则，不妨把在能量积分中增加"点项"，即能量积分中同时含有点、线、面三类项，这样问题(14.1.5~6)稍微推广为如下的形式

$$\begin{cases}J(u)=\displaystyle\iint_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy\\ \qquad\qquad+\displaystyle\int_{\Omega'}\left[\frac{1}{2}\eta u^2-qu\right]ds+\sum_{\Omega''}[-pu]=\text{极小} & (14.3.1)\\ \Omega_0:\ u=\bar{u} & (14.3.2)\end{cases}$$

这里 Ω 就是定解域，它当然可以表为所有面元之和。Ω', Ω'', $\Omega_0\subset\Omega+\partial\Omega$，分别表示需要计算能量的线及点的总和。$\Omega_0$ 表示施以强加条件的线及点元的总和。Ω', Ω'', Ω_0 可以不仅仅局限在边界 $\partial\Omega$ 上，而且可以展至 Ω 的内部。已知系数 β, η, f, q, $\bar{u}$ 分别定义在有关的部位上。问题(14.1.5~6)就是 $\Omega'=\Gamma_0'$, $\Omega''=$空, $\Omega_0=\Gamma_0$ 的特殊情况。

对域 Ω 作三角剖分，并保证这个剖分与 Ω 原有的分割及系数的间断性相协调，于是 $J(u)$ 可以分解为有关单元上能量积分之和

$$J(u)=\sum_{C\in\Omega}J_C(u)+\sum_{B\in\Omega'}J_B(u)+\sum_{A\in\Omega''}J_A(u)$$

$$J_C(u)=\iint_C\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}dx\,dy$$

$$J_B(u)=\iint_B\left[\frac{1}{2}\eta u^2-qu\right]ds$$

$$J_A(u)=\left[\frac{1}{2}\mu u^2-pu\right]_A$$

在三角剖分(14.2.1 节)的基础上，将未知函数 $u(x, y)$ 代以由它在顶点 $A_1, \cdots, A_{N_0}$ 的值 $u_1, \cdots, u_{N_0}$ 产生的分片线性插值函数 $U(x, y)$，而 $J(u)$ 代以 $J(U)$，这一步可以先按单元执行(14.3.1 节)然后累加起来(14.3.2 节)。

14.3.1 单元分析

在各级单元上，u 代以其顶点值的线性插值。由于剖分与 Ω 由系数的间断性相协调，故在每个单元上，系数 β, f, η, q 等是局部光滑的，为简便计，不妨取为常数。

(一)面元分析 $C=(A_1, A_2, A_3)$

$$u \sim U=\sum_1^3 u_i\lambda_i,\quad \frac{\partial U}{\partial x}=\sum_1^3 u_i\frac{\partial\lambda_i}{\partial x},\quad \frac{\partial U}{\partial y}=\sum_1^3 u_i\frac{\partial\lambda_i}{\partial y}$$

$$J_C(u) \sim J_C(U)=\iint_C\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fU\right\}dx\,dy$$

由于

$$U^2=\left(\sum_1^3 u_i\lambda_i\right)\left(\sum_1^3 u_j\lambda_j\right)=\sum_{i,j=1}^3 u_iu_j\lambda_i\lambda_j$$

$$\left(\frac{\partial U}{\partial x}\right)^2=\sum_{i,j=1}^3 u_iu_j\frac{\partial\lambda_i}{\partial x}\frac{\partial\lambda_j}{\partial x},\quad \left(\frac{\partial U}{\partial y}\right)^2=\sum_{i,j=1}^3 u_iu_j\frac{\partial\lambda_i}{\partial y}\frac{\partial\lambda_j}{\partial y}$$

因此

$$J_C=J_C(u_1, u_2, u_3)=\frac{1}{2}\sum_{i,j=1}^3 a_{ij}^{(C)}u_iu_j-\sum_{i=1}^3 b_i^{(C)}u_i$$

$$a_{ij}^{(C)}=\iint_C\left[\beta\frac{\partial\lambda_i}{\partial x}\cdot\frac{\partial\lambda_j}{\partial x}+\beta\frac{\partial\lambda_i}{\partial y}\cdot\frac{\partial\lambda_j}{\partial y}+\gamma\lambda_i\lambda_j\right]dx\,dy$$

$$b_i^{(C)}=\iint_C f\lambda_i\,dx\,dy$$

系数 β, γ, f 在 C 上离散成为常数，于是根据表 14.1 得到

$$a_{ij}^{(C)}=(\beta\eta_i\eta_j+\beta\xi_i\xi_j)/2D_0=a_{ji}^{(C)}$$

$$b_i^{(C)}=fD_0/b$$

对于一些更复杂的问题如(14.1.18)，面元积分为

$$J_C=\iint_C\left\{\frac{1}{2}\left[\beta_{11}\left(\frac{\partial u}{\partial x}\right)^2+\beta_{12}\left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial u}{\partial y}\right)+\beta_{21}\left(\frac{\partial u}{\partial y}\frac{\partial u}{\partial x}\right)\right.\right.$$
$$\left.\left.+\beta_{22}\left(\frac{\partial u}{\partial y}\right)^2+\gamma u^2\right]-fu\right\}dx\,dy$$

$$\beta_{ij}=\beta_{ji}$$

则用类似的方法可以得到

$$a_{ij}^{(C)}=a_{ji}^{(C)}=\iint_C\left[\beta_{11}\frac{\partial\lambda_i}{\partial x}\frac{\partial\lambda_j}{\partial y}+\beta_{12}\frac{\partial\lambda_i}{\partial x}\frac{\partial\lambda_j}{\partial y}+\beta_{21}\frac{\partial\lambda_i}{\partial y}\frac{\partial\lambda_j}{\partial x}\right.$$
$$\left.+\beta_{22}\frac{\partial\lambda_i}{\partial x}\frac{\partial\lambda_j}{\partial y}+\gamma\lambda_i\lambda_j\right]dx\,dy$$
$$=(\beta_{11}\eta_i\eta_j-\beta_{12}\eta_i\xi_j-\beta_{21}\xi_i\eta_j+\beta_{22}\xi_i\xi_j)/\partial D_0+\gamma D_0(1+\delta_{ij})/24$$

$b_i^{(C)}$ 同前。

当 $C=(A_{n_1}, A_{n_2}, A_{n_3})$ 即顶点标号为 (n_1, n_2, n_3) 时，计算 a_{ij}, b_i 的公式中用到的坐标 x_i, y_i 应取为 x_{n_i}. y_{n_i}, 而能量表达式中 u_i 应代为 u_{n_i}, $i=1, 2, 3$, 因此

$$J_C=J_C(u_{n_1},\ u_{n_2},\ u_{n_3})=\frac{1}{2}\sum_{i,j=1}^{3}a_{ij}^{(C)}u_{n_i}u_{n_j}-\sum_{i=1}^{3}b_i^{(C)}u_{n_i}\tag{14.3.3}$$

应该指出，对于介质系数如 β, … 采用分单元的离散化方法事实上就是对于介质间断性的一种自动处理，不论这种间断性在几何上复杂到什么程度，只要剖分是协调的，即把介质间断的点、线元落在剖分的点、线元上，就自动体现了交界条件(14.1.17)。

(二)线元分析 $B=(A_1,\ A_2)$

$$u\sim U=\sum_1^2 u_i\lambda_i,\quad \frac{\partial U}{\partial s}=\sum_1^2 u_i\frac{\partial\lambda_i}{\partial s}$$

$$J_B(u)\sim J_B(U)=\int_B\left[\frac{1}{2}\eta U^2-qU\right]ds$$

于是

$$J_B=J_B(u_1,\ u_2)=\frac{1}{2}\sum_{i,j=1}^{2}a_{ij}^{(B)}u_iu_j-\sum_{i=1}^{2}b_i^{(B)}u_i$$

$$a_{ij}^{(B)}=\int_B\eta\lambda_i\lambda_j ds=a_{ji}^{(B)}$$

$$b_i^{(B)}=\int_B q\lambda_i ds$$

系数 η, q 在 B 上取常数值，故由表 14.3 得

$$a_{ij}^{(B)}=\eta L(1+\delta_{ij})/6$$
$$b_i^{(B)}=qL/2$$

在有些问题，例如本身是一维问题如(14.1.20)，或者具有复杂结构的问题中，能量中的线项取如下的形式

$$J_B=\int_B\left\{\frac{1}{2}\left[\xi\left(\frac{\partial u}{\partial s}\right)^2+\eta u^2\right]-qu\right\}ds=\frac{1}{2}\sum_{i,j=1}^{2}a_{ij}^{(B)}u_iu_j-\sum_{i=1}^{2}b_i^{(B)}u_i$$

这时

$$a_{ij}^{(B)}=\int_B\left[\xi\frac{\partial\lambda_i}{\partial s}\frac{\partial\lambda_j}{\partial s}+\eta\lambda_i\lambda_j\right]ds=(-1)^{i+j}\xi/L+\eta L(1+\delta_{ij})/6=a_{ji}^{(B)}$$

$b_i^{(B)}$ 同前。

当 $B=(A_{m_1},\ A_{m_2})$ 即顶点标号为 $(m_1,\ m_2)$ 时，计算时用到的坐标 x_i, y_i 应取为 x_{m_i}, y_{m_i}，而 u_i 代为 u_{m_i}, $i=1,\ 2$，因此

$$J_B=J_B(u_{m_1},\ u_{m_2})=\frac{1}{2}\sum_{i,j=1}^{2}a_{ij}^{(B)}u_{m_i}u_{m_j}-\sum_{i=1}^{2}b_i^{(B)}u_{m_i}\tag{14.3.4}$$

(三)点元分析 $A=(A_1)$

在点元 $A=(A_1)$ 上离散化是很显然的。事实上，在顶点 A_1 处 $u=u_1$, u 的线性插值也有 $U=u_1$; 同时，能量"积分"已经是离散形式的。因此

$$J_A(u)=[-pu]_A=J_A(U)=J_A(u_1)=\frac{1}{2}a_{11}^{(A)}u_1u_1-b_1^{(A)}u_1$$

这里矩阵 $a_{ij}^{(A)}$, $b_i^{(A)}$ 为一阶的，均退化为一个数

$$a_{11}^{(A)}=0$$
$$b_1^{(A)}=p$$

当问题本身是一维问题，或者在复杂结构的问题中，能量中的点项作如下形式

$$J_A=\left[\frac{1}{2}\mu u^2-pu\right]_A=\frac{1}{2}a_{11}^{(A)}u_1u_1-b_1^{(A)}u_1$$

这时

$$a_{11}^{(A)}=\mu$$
$$b_1^{(A)}=p$$

当 $A=A_l$ 即该点元的标号为(l)时

$$J_A=J_A(u_l)=\frac{1}{2}a_{11}^{(A)}u_lu_l-b_1^{(A)}u_l \tag{14.3.5}$$

14.3.2 总体合成

能量积分分单元离散化后，总体的能量就成为

$$J(u)\sim J(U)=J(u_1,\ \cdots,\ u_{N_0})=\sum_{C\in\Omega}J_C+\sum_{B\in\Omega'}J_B+\sum_{A\in\Omega''}J_A$$
$$=\frac{1}{2}\sum_{i,j=1}^{N_0}a_{ij}u_iu_j-\sum_{i=1}^{N_0}b_iu_i$$

它的系数矩阵 $\boldsymbol{A}=(a_{ij})$，$\boldsymbol{b}=(b_i)$可由各单元的系数阵 $a_{ij}^{(C)}$，$b_i^{(C)}$，…(14.3.1 节)以适当的方式累加而得。

为此，在解题开始应具备下列有关剖分的几何量及物理量的信息：

1．点元的标号和坐标：$(x_k,\ y_k)$，$k=1,\ \cdots,\ N_0$。

2．Ω 中的面元的三顶点标号$(n_{1_k},\ n_{2_k},\ n_{3_k})$和相应的系数 β_k，f_k；$k=1,\ \cdots,\ N_2$。

3．Ω' 中的面元的两顶点标号$(m_{1_k},\ m_{2_k})$和相应的系数 η_k，q_k；$k=1,\ \cdots,\ M_1$。

4．Ω'' 中的点元标号(l_k)和相应系数 p_k；$k=1,\ \cdots,\ M_0$。

5．Ω_0 中的点元标号(h_k)和该点的强加值 $\bar{u}_k$；$k=1,\ \cdots,\ M$ (见 14.3.3 节)。

此外，应根据 14.3.1 节编出三个标准化的面、线、点单元分析程序，它们能从单元的几何及物理信息产生单元系数阵。

在这个基础上，总体系数阵的合成过程如下：

(1) 首先对待定阵 $\boldsymbol{A}$，$\boldsymbol{b}$ 的全部元素置 0，即

$$0\Rightarrow a_{ij},\ 0\Rightarrow b_i,\quad i,\ j=1,\ \cdots,\ N_0$$

(2) 对 $C\in\Omega$ 逐个作面元分析，也就是根据单元顶点序号 n_i 和坐标 x_{n_i}，$y_{n_i}(i=1,\ 2,\ 3)$ 和参数 β，γ，f 算出单元系数 $a_{ij}^{(C)}$，$b_i^{(C)}$ 并根据(14.3.3) 把它们分别累加到总体阵 $\boldsymbol{A}$，$\boldsymbol{b}$ 的适当部位，即

$$a_{ij}^{(C)}+a_{n_in_j}\Rightarrow a_{n_in_j},\ b_i^{(C)}+b_{n_i}\Rightarrow b_{n_i},\quad i,\ j=1,\ 2,\ 3$$

注意这是在既有基础上的累加而不是取代，不同的单元可以对于同一位置的系数都有贡献。

(3) 对 $B\in\Omega'$ 逐个作线元分析。由顶点序 m_i 和坐标 x_{m_i}，y_{m_i} 和参数 ξ，η，q 算出线元系数阵 $a_{ij}^{(B)}$，$b_i^{(B)}$ 再按照(14.3.4)累加

$$a_{ij}^{(B)}+a_{m_im_j}\Rightarrow a_{m_im_j},\ b_i^{(B)}+b_{m_i}\Rightarrow b_{m_i},\quad i,\ j=1,\ 2$$

(4) 对点元 $A\in\Omega''$ 逐个作点元分析，根据顶点序号 l 及 μ，p 算出 $a_{11}^{(A)}$，$b_1^{(A)}$ 再按(14.3.5)累加

$$a_{11}^{(A)}+a_{ll}\Rightarrow a_{ll},\ b_1^{(A)}+b_l\Rightarrow b_l$$

全部单元处理完毕后就得到总体矩阵 $\boldsymbol{A}$，$\boldsymbol{b}$，由于各个单元矩阵 $a_{ij}^{(C)}$，$a_{ij}^{(B)}$，$a_{ij}^{(A)}$ 是对称的，所以总体阵 $\boldsymbol{A}$ 也是对称的，

$$a_{ij}=a_{ji},\quad i,\ j=1,\ \cdots,\ N_0$$

这样能量积分即二次泛函 $J(u)$ 就完全离散化成为多元二次函数

$$J(u_1,\ \cdots,\ u_{N_0})=\frac{1}{2}\sum_{i,j=1}^{N_0} a_{ij}u_iu_j-\sum_{i=1}^{N_0}b_iu_i \tag{14.3.6}$$

首先考虑没有强加条件的情况，即 $\Omega_0=0$(空集)。这时原问题(14.3.1～2)成为无条件变分问题

$$\text{在函数类}\ S\ \text{内定}\ u\ \text{使得}\quad J(u)=\text{极小} \tag{14.3.7}$$

这里 S 是所有不受强加约束的，具有一定光滑性使得积分 $J(u)$ 有意义的函数类，它有无穷多自由度。离散化后，(14.3.7)变成

$$\text{在函数类}\ S'\ \text{内定}\ u\ \text{使得}\quad J(u)=\text{极小} \tag{14.3.8}$$

这里 S' 是所有片状线性插值函数所组成的函数类，是 S 的一个子类，$S'\subset S$，S' 只有有限多自由度，即有 N_0 个自由参数 $u_1,\ v_2,\ \cdots,\ u_{N_0}$。问题(14.3.8)就是多元二次函数(14.3.6)的无条件极小问题即

$$\text{定参数}\ u_1,\ \cdots,\ u_{N_0}\ \text{使得}\quad J(u_1,\ \cdots,\ u_{N_0})=\text{极小} \tag{14.3.9}$$

根据微积分中的极值原理(14.1.13～15)，当二阶导数阵 $\dfrac{\partial^2 J}{\partial u_i\partial u_j}$ 正定时，极小问题(14.3.9)等价于解方程组

$$\frac{\partial J}{\partial u_i}=0,\quad i=1,\ \cdots,\ N_0 \tag{14.3 10}$$

由于(14.3.6)，J 是二次的，它的一阶导数是一次的，即

$$\frac{\partial J}{\partial u_1}=a_{11}u_1+\frac{1}{2}(a_{12}+a_{21})u_2+\frac{1}{2}(a_{13}+a_{31})u_3+\cdots-b_1=\sum_{j=1}^{N_0}a_{ij}u_j-b_1$$

这里利用了对称性 $a_{ij}=a_{ji}$。一般地有

$$\frac{\partial J}{\partial u_i}=\sum_{j=1}^{N_0}a_{ij}u_j-b_i=0,\quad i=1,\ \cdots,\ N_0$$

J 的二阶导数则都是常数，与 $u_1,\ \cdots,\ u_{N_0}$ 无关:

$$\frac{\partial^2 J}{\partial u_i\partial u_j}=a_{ij}$$

因此，当系数阵 $\boldsymbol{A}=[a_{ij}]$ 为对称正定时，$(u_1,\ \cdots,\ u_{N_0})$ 使二次函数 J 达到极小的充要条件是满足线代数方程组

$$\sum_{j=1}^{N_0}a_{ij}u_i=b_i,\quad i=1,\ \cdots,\ N_0 \tag{14.3.11}$$

即

$$\boldsymbol{Au}=\boldsymbol{b} \tag{14.3.12}$$

因此变分问题就最终离散化为解线性代数方程组(14.3.11)的问题，注意方程组的系数阵 $\boldsymbol{A}$, $\boldsymbol{b}$ 就是能量函数(14.3.6)中的二次及一次部分的系数阵。

有时，在定出能量函数时可能得出不对称的系数阵 $\boldsymbol{A}=[a_{ij}]$。由于二次型的系数总可以对称化而型值不变

$$\sum_{i,j=1}^{N_0}a_{ij}u_iu_j\equiv\sum_{i,j=1}^{N_0}\frac{1}{2}(a_{ij}+a_{ji})u_iu_j$$

那末，极小化的代数方程组的系数阵就不是 $\boldsymbol{A}=[a_{ij}]$ 而是它的对称化

$$\frac{1}{2}(\boldsymbol{A}+\boldsymbol{A}^T)=\left[\frac{1}{2}(a_{ij}+a_{ji})\right]$$

也就是必需经过对称化才能得到正确的代数方程组的系数阵。如果在单元分析的一级上单元系数阵——"小"矩阵——是对称的(14.3.1节中就是这样)或进行了对称化,则就能保证总体系数阵——"大"矩阵——的对称性。

由于 J 是二次的,它的二阶以上的偏导数均为零,故有

$$J(u_1+\delta u_1,\ \cdots,\ u_{N_0}+\delta u_{N_0})=J(u_1,\ \cdots,\ u_n)+\delta J(u_1,\ \cdots,\ u_{N_0};\ \delta u_1,\ \cdots)$$

$$J(u_1+\delta u_1,\ \cdots,\ u_{N_0}+\delta u_{N_0})=J(u_1,\ \cdots,\ u_{N_0})+\delta J(u_1,\ \cdots,\ u_{N_0};\ \delta u_1,\ \cdots,\ \delta u_{N_0})+\frac{1}{2}\delta^2 J(\delta u_1,\ \cdots,\ \delta u_{N_0}) \tag{14.3.13}$$

这 δJ, $\delta^2 J$ 就是函数 J 在点 $(u_1,\ \cdots,\ u_{N_0})$ 的一次及二次微分

$$\delta J(u_1,\ \cdots,\ u_{N_0};\ \delta u_1,\ \cdots,\ \delta u_{N_0})=\sum_{i=1}^{N_0}\frac{\partial J}{\partial u_i}\delta u_i=\sum_{i=1}^{N_0}\left(\sum_{j=1}^{N_0}a_{ij}u_j-b_i\right)\delta u_i \tag{14.3.14}$$

$$\delta^2 J(\delta u_1,\ \cdots,\ \delta u_{N_0})=\sum_{i,j=1}^{N_0}\frac{\partial^2 J}{\partial u_i\partial u_j}\delta u_i\delta u_j=\sum_{i,j=1}^{N_0}a_{ij}\delta u_i\delta u_j \tag{14.3.15}$$

从(14.3.13)中各项的量级对比也可以看出,即使二阶导数阵 $\frac{\partial^2 J}{\partial u_i\partial u_j}=a_{ij}$ 仅仅是半正定,也就是说

$$\frac{1}{2}\sum_{i,j=1}^{N_0}a_{ij}\delta u_i\delta u_j\geqslant 0,\quad 对一切\ \delta u_i,\ i=1,\ \cdots,\ N_0$$

时,极小问题(14.3.9)也等价于解方程组(14.3.10)即(14.3.11)。

泛函 $J(u)$ (14.3.1)的二次变分是(参考(14.1.10))

$$\frac{1}{2}\delta^2 J(\delta u)=\iint_{\Omega}\frac{1}{2}\beta\left(\frac{\partial\delta u}{\partial x}\right)^2+\beta\left(\frac{\partial\delta u}{\partial y}\right)^2 dx\,dy+\int_{\Omega'}\frac{1}{2}\eta(\delta u)^2 ds \tag{14.3.16}$$

如果命 δu 为由 $\delta u_1,\ \cdots,\ \delta u_{N_0}$ 产生的分片线性插值函数,$\delta u\in S'$,则由14.3.1节,14.3.2节的离散化方法不难看出,作为函数 $J(u_1,\ \cdots,\ u_n)$ 的二次微分(14.3.15)与作为泛函 $J(u)$ 的二次变分是一致的,即

$$\frac{1}{2}\delta^2 J(\delta u)\equiv\frac{1}{2}\delta^2 J(\delta u_1,\ \cdots,\ \delta u_{N_0})\equiv\frac{1}{2}\sum_{i,j=1}^{N_0}a_{ij}\delta u_i\delta u_j \tag{14.3.17}$$

设 $\beta>0$, $\eta\geqslant 0$ 并且 $\eta\not\equiv 0$。在此情况下在14.1.2节中已证明了二次变分(14.3.16)对于函数类 $S=S_0$ 的正定性,因此在其子类 $S'\subset S$ 上当然还是正定的,因此二次型(14.3.17)正定,即矩阵 $\boldsymbol{A}=[a_{ij}]$ 正定,从而保证线代数方程(14.3.11)有唯一解。

设 $\beta>0$, $\eta\equiv 0$,这就是所谓第二类边值问。用14.1.2中的方法可知二次变分对于 S 为退化半正定,而且

$$\delta^2 J(\delta u)=0\Leftrightarrow\delta u\equiv c=常数 \tag{14.3.18}$$

由于 $\delta u\equiv c\in S_0$,即相当于用 $\delta u_1=\cdots=\delta u_{N_0}=c$ 插出的分片线性函数,所以二次型(14.3.17)也是退化半正定而且

$$\frac{1}{2}\sum_{i,j=1}^{N_0}a_{ij}\delta u_i\delta u_j=0\Leftrightarrow\delta u_1=\cdots=\delta u_{N_0}=c \tag{14.3.19}$$

因此矩阵 $\boldsymbol{A}$ 是退化半正定,行列式 $|\boldsymbol{A}|=0$。

对此退化的情况,按照14.1.2节所述:

1. 齐次问题——即在(14.3.1)中命 $f\equiv 0$, $q\equiv 0$, $p\equiv 0$——有非零解

$$u\equiv 1 \tag{14.3.20}$$

而任意非零解可以表为这个解的常数倍即 $u\equiv c$。由于这些非零解都含在子函数类 S' 中，因此离散后对应于(14.3.9)的齐次问题有同样的非零解即

$$u_1=\cdots=u_{N_0}=1 \tag{14.3.21}$$

而任意的非零解可表为它的常数倍，即 $u_1=\cdots=u_{N_0}=c$。注意离散的齐次问题(14.3.9)就是齐次线代数方程

$$\sum_{j=1}^{N_0} a_{ij}u_j=0,\quad i=1,\ \cdots,\ N_0 \tag{14.3.22}$$

2. 非齐次问题(14.3.8)有解的充要条件即所谓协调条件是

$$\iint_\Omega f\,dx\,dy+\int_{\Omega'} q\,ds+\sum_{\Omega''} p=0 \tag{14.3.23}$$

这在物理上相当于外载荷的平衡条件，是(14.1.30)的推广。当有解时，任意两个解必相差一个常数即相差一个齐次问题的解。在离散化得到退化、对称的线代数方程组(14.3.11)，根据线代数的初等理论，(14.3.11)有解的充要条件是右项向量 $\boldsymbol{b}=(b_1,\ \cdots,\ b_{N_0})$ 与齐次方程组(14.3.22)的基本解向量——现在就是(14.3.21)——正交，即

$$\sum_{i=1}^{N_0} b_i=0 \tag{14.3.24}$$

这就是代数方程组(14.3.11)的协调条件。根据 14.3.1 节，14.3.2 节的分析方法可知 b_i 是由系数 $f,\ q,\ p$ 经离散化而得来的。问题在于：当原给的系数 $f,\ q,\ p$ 满足协调条件(14.3.23)时，经过离散化后是否自动保证(14.3.24)成立？答案是肯定的。如果在单元分析中涉及 $f,\ q,\ p$ 的积分是准确的，即没有作任何近似，则可以证明

$$\sum_{i=1}^{N_0} b_i=\iint_\Omega f\,dx\,dy+\int_{\Omega'} q\,ds+\sum_{\Omega''} p$$

事实上只须取 $u\equiv1$ 即 $u_i\equiv1$，于是根据(14.3.1)，(14.3.6)，(14.3.19)即得

$$J(u)=-\int_\Omega f\,dx\,dy-\int_{\Omega'} q\,ds-\sum_{\Omega''} p=J(u_1,\ \cdots,\ u_{N_0})=-\sum_{i=1}^{N_0} b_iu_i$$

因此原问题的协调条件(14.3.23)自动保证了离散问题(14.3.11)的协调条件(14.3.24)，从而保证离散问题有解，而任意两个解向量的差是一个常向量。在实践中，往往对 $f,\ q,\ p$ 作近似的处理，例如取为分片常数，于是离散化后有可能不严格满足(14.3.24)，例如

$$\sum_{i=1}^{N_0} b_i=\varepsilon\neq0$$

对此可将 b_i 稍修改，即

$$b_i-\frac{\varepsilon}{N_0}\Rightarrow b_i,\quad i=1,\ \cdots,\ N_0$$

这样新的右端满足协调条件(14.3.24)，保证退化方程组有解(见第十三章§13.3之末段)。

上面的例子说明了，原变分问题的正定性或退化半正定性以及解的唯一性或多重性结构，经有限元离散化后，一般能得到忠实地保持，这是有限元法的一个优点。

当原问题为正定或半正定时，离散方程组的解点就是能量函数 $J(u_1,\ \cdots,\ u_{N_0})$ 的极小点。当原始变分问题为不定时，通常上所要求的只是能量 J 达到临界，在离散化后，系数阵 $\boldsymbol{A}$ 也是不定的，但所要求的只是能量函数 $J(u_1,\ \cdots,\ u_{N_0})$ 达到临界，即 $\dfrac{\partial J}{\partial u_i}=0$，这时待解的方程组仍然是(14.3.11)，不过它解点不一定是 J 的极小点而已。

14.3.3 强加条件和缝隙的处理

在有强加条件的情况下，还需要对上面得到的能量系数阵 $\boldsymbol{A}$, $\boldsymbol{b}$ 作适当的处理后才能得到最终定解的代数方程组。

施以强加条件(14.3.2)的集合 Ω_0 是一些线元及点元的组合。在每个线元上 $\bar{u}(s)$ 可以离散化为其两个顶点(设为 A_1, A_2)的值 $\bar{u}_1$, $\bar{u}_2$ 的线性插值。这和在 Ω 上采取的分片线性插值法是协调的。因此只须对 Ω_0 内所有顶点，命其序号为 h_1, …, h_M 规定条件

$$u_{h_k}=\bar{u}_k, \quad k=1, \cdots, M \tag{14.3.25}$$

于是变分问题(14.3.1～2)就离散化为二次函数 $J(u_1, \cdots, u_{N_0})$ 在条件(14.3.25)下的极值问题。注意 J 中一部分变量取已知值，因此可以视 J 为其余变量的函数，而原来的条件极值问题就成为对于其余变量的无条件极值问题，即满足线方程组

$$\frac{\partial J}{\partial u_i}=0, \quad i=1, \cdots, N_0, \quad i\neq h_1, \cdots, h_M \tag{14.3.26}$$

事实上只须从方程组(14.3.10)中删去 $i=h_1$, …, h_M 的 M 个方程，而在余下的方程中代进已知值(14.3.25)相应项移到右端，故得到 N_0-M 个方程(系数仍然是对称的)和相同个数的未知数。

另一个等价的办法对原有矩阵 $\boldsymbol{A}$, $\boldsymbol{b}$ 作下列形式的修改

$$b_i \text{ 修改为} \begin{cases} \bar{u}_i, & \text{当 } i=h_1, \cdots, h_M \\ b_i-\sum\limits_{k=1}^{M} a_{ih_k}\bar{u}_k, & \text{当 } i\neq h_1, \cdots, h_M \end{cases}$$

$$a_{ij} \text{ 修改为} \begin{cases} 0, & \text{当 } i\neq j,\ i \text{ 或 } j=h_1, \cdots, h_M \\ 1, & \text{当 } i=j,\ i=h_1, \cdots, h_M \end{cases}$$

这样仍为 N_0 个方程(系数也对称)和 N_0 个未知数，在程序实现中可以避免由于删去方程和未知数而引起的重新编号的麻烦。

有时强加条件经离散化后不象(14.3.25)那样简单而是取如下更一般的形式

$$\sum_{j=1}^{N_0} c_{ij}u_j=d_i, \quad i=1, \cdots, M \tag{14.3.27}$$

要求在这个约束条件定函数 $J(u_1, \cdots, u_{N_0})$ 的极值。对此可以采用所谓拉格朗日乘子法。引进新的变量 λ_1, …, λ_M, 作二次函数

$$G(u_1, \cdots, u_{N_0}, \lambda_1, \cdots, \lambda_M)=J(u_1, \cdots, u_{N_0})+\sum_{i=1}^{M}\lambda_i\left(\sum_{j=1}^{N_0}c_{ij}u_j-d_i\right)$$

它有 N_0+M 个变量。可以证明，J 的条件极值问题等价于 G 的无条件极值问题。后者的极值条件是

$$\begin{cases} \dfrac{\partial G}{\partial u_i}=0, & i=1, \cdots, N_0 \\ \dfrac{\partial G}{\partial \lambda_i}=0, & i=1, \cdots, M \end{cases}$$

这就是

$$\begin{aligned} &\sum_{j=1}^{N_0}a_{ij}u_j+\sum_{j=1}^{M}c_{ji}\lambda_j=b_i, \quad i=1, \cdots, N_0 \\ &\sum_{j=1}^{N_0}c_{ij}u_j=d_i, \quad i=1, \cdots, M \end{aligned} \tag{14.3.28}$$

也可表为矩阵形式

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}^T \\ \boldsymbol{C} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix} \tag{14.3.29}$$

这个方法在形式上比较简单,新的系数矩阵保持了对称性。但是,方程组的阶数从 N_0 扩大到 N_0+M, 正定性则不一定能保持。

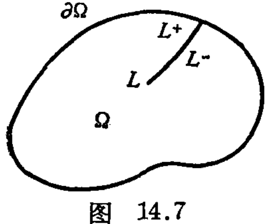


图 14.7

缝隙的处理

如图 14.7 所示有时定解区域 Ω 内有缝隙,缝隙在物理上总是有宽度的,但当宽度相对地小时,可以视为无宽度的曲线 L, 有正负两岸 L^+, L^-。设想在问题中规定缝隙的两岸都是自由边,这时泛函(14.3.1)的形式不变,但 $u(x, y)$在两岸互相独立,可以取不同的值。当能量达到极小时,在缝隙两岸自动满足自由边界条件

$$\left(\beta\frac{\partial u}{\partial \nu}\right)^+=0, \quad \left(\beta\frac{\partial u}{\partial \nu}\right)^-=0$$

对于这种情况,在作剖分时应使缝 L 落在线元上,缝上的每个点元和线元都一分为二,变为双重点、双重线,各有自己的编号,分别赋有 u 值,分属于正负两岸,相同的仅仅是它们的位置坐标。与缝隙相邻接的面元自然是分属两岸,应该视为分离的面元,不再具有公共边。在解题时,事实上只需在初始阶段,即进行剖分和准备相应信息时照上述原则办理,在以后的阶段就无须再作特殊处理。

14.3.4 代数计算和结果解释

有限元法离散化后得到线代数组 $\boldsymbol{Au}=\boldsymbol{b}$ 的系数阵 $\boldsymbol{A}$ 总是正定的。当原始变分问题具有正定性时,在有限元法中一般也保证系数阵 $\boldsymbol{A}$ 的正定性,除了由于强加条件的处理有时会有些麻烦。这种对称正定性的特点保持,是有限元法的一个优点。由于对称正定阵的计算方法发展得比较完善,因此这种特性保持对于方程的解算也是有利的。

有限元法所得的系数阵的另一特点是稀疏性,即绝大多数的元素为零,也就是说组中每个方程中只有少数几个特定部位的系数不为 0,即矩阵基本上是带状的。事实上,从 14.3.1～3.2 节中可以看出,每当点元 A_i 与 A_j 不相邻,即不同属于某个线元或面元时,在合成后的能量函数中就不出现 u_iu_j 的项,即相应的矩阵元素 a_{ij} 必为零。这种稀疏性对于实际解算提供了有利条件,在程序上可采取压缩零的技巧使得仅仅 $\boldsymbol{A}$ 中的非零元素才被存储,可以节约存储量和运算量。

针对着系数阵 $\boldsymbol{A}$ 的对称正定性和稀疏性,在实际解算时可以用超松弛法、分块超松弛法或其它类似的迭代法。也可以采用如分块消元法或其它适合稀疏块状结构的直接法。还可以采用迭代法和直接法相结合的共轭斜量法,这也是适合与对称正定和稀疏特点的。具体算法可以参考第九章。

代数计算结束后就得到解在离散点 $A_1, \cdots, A_{N_0}$ 的值 $u_1, \cdots, u_{N_0}$。在实践上常常需要知道导数 $\beta\frac{\partial u}{\partial x}$, $\beta\frac{\partial u}{\partial y}$ 的分布以及 u 在其它点的值。因此需要再作一轮单元上的结果分析,即按照原来的插值原则补插算出所需要的量。例如对于导数,在每个面元 $C=(A_1, A_2, A_3)$ 上按照(14.2.7)取为

$$\beta\frac{\partial u}{\partial x}=\beta(\eta_1u_1+\eta_2u_2+\eta_3u_3)/D$$

$$\beta\frac{\partial u}{\partial y}=-\beta(\xi_1u_1+\xi_2u_2+\xi_3u_3)/D$$

并作为在单元中点的值。当需要知道在节点处的导数值时则可以取相邻面元中点值的适当的平均值。特别是利用计算机来显示或制作等值线图或向量场图或其它曲线是非常有利于结果分析的。

14.3.5 方法的特点

有限元法的特点以及与其它方法的对比可以综述如下:

(1) 有限元法是以变分原理和剖分插值为基础的。它把在无穷多自由度的函数类S中的极值问题代为S的一个有限多自由度的子类S'中的极值问题,在这点上有限元法是传统的能量法(即李兹-加辽金法)的一种变形。在传统的能量法中,子类S'是由解析函数组成的,缺乏灵活性,而有限元法则是在剖分即格网插值的基础上,来形成子类S',在这点上,它是差分法的一种变形,吸取并发扬了后者的灵活性。因此有限元法是能量法与差分法相结合而发展了的方法。

(2) 在有限元法中,最终求解的多次二次函数的极值方程,系数阵总是对称的,而且当原始问题为正定时,离限化后一般也保持正定性。这一特点是能量法中共同的,而在差分法中则不一定总能做到。有限元系数阵又是稀疏的,这一特点是差分法中共同的,但传统的能量法中则不然。对称正定与稀疏特性对于数值解算是有利的。

(3) 有限元法的各个环节,如单元分析、总体合成、代数解算、结果解释等等在程序实现上都是便于标准化的。至少对于同一类型的问题,不论几何形状或物理参数分布如何,不论采用什么插值方法,都可以用同一套标准程序来对付。对于解题者说来,只须准备有关剖分的几何、物理参数的最低限度的信息即可,这样可以大大缩短解题周期。

(4) 在有限元法中,不论问题是简单或复杂,基本上是同等对待的。因此,对于规则区域和常系数的问题而言,有限元法的效率会比一般差分法低,但是,随着问题在几何上物理上的复杂性的增高而优点愈显。有限元法主要是面对这类问题的。

(5) 有限元法利用了变分原理和剖分插值比较成功地解决了自然边界条件的处理问题。但是,强加条件处理上的矛盾则相对地上升,还有待于改进。

(6) 本章没有讨论有限元法的收敛性问题,即当剖分愈来愈细时,离散解是否愈来愈趋近于真解的问题。有限元法的基础理论实际上是相当简单的,在相当广泛的范围内,可以确保在能量积分意义下的收敛性,从而保证方法的可靠性。这也是有限元法的一个特点,见[2]。

§14.4 有限元法的一些应用

有限元法对于椭圆型问题是普遍适用的。在§14.3中通过平面二阶椭圆方程边值问题的典型例子介绍了基本方法。本节再介绍在几何上,解析上、物理上有些特点的问题,如轴对称、本征值和平面弹性问题,仍用三角形线性插值。最后介绍提高精度的三角形二次插值。关于其它的剖分和插值方法,三维问题,涉及四阶椭圆方程的板、壳问题,以及含有时间的动态问题等等则可以参考专门的著作如[3]。

14.4.1 轴对称问题

平面椭圆方程的变分原理(14.1.18)自然地推广到空间(见 §14.1),变分问题

$$\begin{cases} J(u)=\iiint\limits_{\bar{\Omega}}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2+\beta\left(\frac{\partial u}{\partial z}\right)^2\right]-fu\right\}dx\,dy\,dz \\ \qquad +\int_{\bar{\Gamma}'_0}\left\{\frac{1}{2}\eta u^2-qu\right\}d\sigma=\text{极小} \\ \bar{\Gamma}_0:\ u=\bar{u} \end{cases}$$

等价于边值问题

$$\begin{cases} \bar{\Omega}-\bar{L}:\ -\left(\frac{\partial}{\partial x}\beta\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial u}{\partial y}+\frac{\partial}{\partial z}\beta\frac{\partial u}{\partial z}\right)=f \\ \bar{L}:\ \left(\beta\frac{\partial u}{\partial \nu}\right)^-=\left(\beta\frac{\partial u}{\partial \nu}\right)^+ \\ \bar{\Gamma}'_0:\ \beta\frac{\partial u}{\partial \nu}+\eta u=q \\ \bar{\Gamma}_0:\ u=\bar{u} \end{cases}$$

当问题具有轴对称性，即区域 $\bar{\Omega}$ 及其内外界面 $\bar{L}$, $\bar{\Gamma}'_0$, $\bar{\Gamma}_0$ 都是回转体或回转面，所有的系数 β, η, f, q 都具有回转不变性时，则解也必具有回转不变性即轴对称性。这时以采取柱坐标 r, z, φ 为便，而且一切量与 φ 无关。可以取一个 $\varphi=0$ 的参考平面(r, z)，其上有二维域 Ω 以及界线 $\partial\Omega$, Γ_0, Γ'_0, L 等等，它们绕 z 轴旋转而生成 $\bar{\Omega}$, $\partial\bar{\Omega}$, $\bar{\Gamma}_0$, $\bar{\Gamma}'_0$, $\bar{L}$。由于

$$\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2+\left(\frac{\partial u}{\partial z}\right)^2=\left(\frac{\partial u}{\partial r}\right)^2+\left(\frac{\partial u}{\partial z}\right)^2+\left(\frac{1}{r}\frac{\partial u}{\partial \varphi}\right)^2=\left(\frac{\partial u}{\partial r}\right)^2+\left(\frac{\partial u}{\partial z}\right)^2$$

$$\iiint\limits_{\bar{\Omega}}\cdots dx\,dy\,dz=\iiint\limits_{\bar{\Omega}}\cdots r\,dr\,dz\,d\varphi=2\pi\iint\limits_{\Omega}\cdots r\,dr\,dz$$

$$\iint\limits_{\bar{\Gamma}}\cdots d\sigma=\iint\limits_{\bar{\Gamma}}\cdots r\,ds\,d\varphi=2\pi\int_{\Gamma}\cdots r\,ds$$

因此三维变分问题可以表为二维的形式

$$\begin{cases} J(u)=2\pi\iint\limits_{\Omega}\left\{\frac{1}{2}\beta\left[\left(\frac{\partial u}{\partial r}\right)^2+\left(\frac{\partial u}{\partial z}\right)^2\right]-fu\right\}r\,dr\,dz \\ \qquad +2\pi\int_{\Gamma'_0}\left[\frac{1}{2}\eta u^2-qu\right]r\,ds=\text{极小} \\ \Gamma_0:\ u=\bar{u} \end{cases} \tag{14.4.1}$$

它等价于边值问题

$$\begin{cases} \Omega:\ -\left(\frac{1}{r}\frac{\partial}{\partial r}r\beta\frac{\partial u}{\partial r}+\frac{\partial}{\partial z}\beta\frac{\partial u}{\partial z}\right)=f \\ L:\ \left(\beta\frac{\partial u}{\partial \nu}\right)^-=\beta\left(\frac{\partial u}{\partial \nu}\right)^+ \\ \Gamma'_0:\ \beta\frac{\partial u}{\partial \nu}+\eta u=q \\ \Gamma_0:\ u=\bar{u} \end{cases} \tag{14.4.2}$$

当 $\beta\equiv1$, $f\equiv0$ 就得到

$$\frac{1}{r}\frac{\partial}{\partial r}r\frac{\partial u}{\partial r}+\frac{\partial^2 u}{\partial z^2}=0 \tag{14.4.3}$$

或

$$\frac{\partial^2 u}{\partial r^2}+\frac{\partial^2 u}{\partial z^2}+\frac{1}{r}\frac{\partial u}{\partial r}=0 \tag{14.4.4}$$

这就是轴对称下的拉普拉斯方程。

域Ω总是位于参考平面即rz平面的右半$r\geqslant 0$，它的边线$\partial\Omega$可能不与z轴接触(如图14.8)但也能有一部分Γ_z在对称轴上，另一部分Γ不在对称轴上(如图14.9)：

$$\partial\Omega=\Gamma_z+\Gamma$$

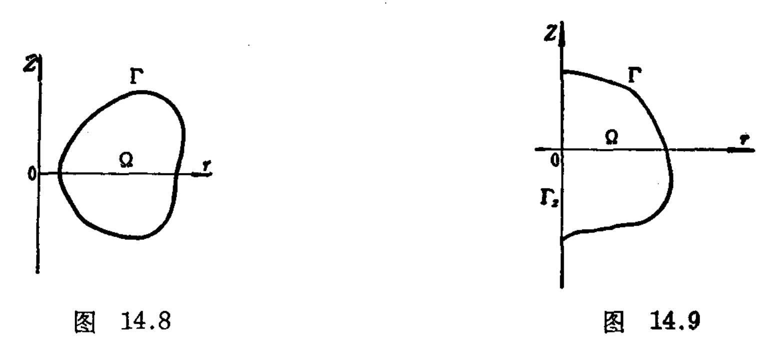


图 14.8　　　　图 14.9

应该记住，问题本来的定解域是由Ω生成的回转体$\bar{\Omega}$，它的边界面$\partial\bar{\Omega}$仅仅是由Γ旋转生成的，与Γ_z无关。Γ_z在对称轴上，实际上并不构成问题的边界。因此在Γ_z上无边界条件可言，边界条件仅作用于Γ，后者照例又可分为强加的Γ_0和自然的Γ_0'两个部分：

$$\Gamma=\Gamma_0+\Gamma_0'$$

从(14.4.2)(14.4.4)看，由于系数含有因子$\frac{1}{r}$，方程在Γ_z上即$r=0$处有奇异性。这在形式化的差分方法中会碰到困难。但是，这只是坐标系的奇异性，$\bar{\Gamma}_z$在回转体$\bar{\Omega}$的内部，作为三维问题本身在那里并没有奇异性。当以变分原理(或守恒原理)为基础来离散化时，上述困难自然地不出现。这种对于不同坐标系统的适应性也是有限元法的一个优点。

对于平面问题的有限元方法只须稍作修改便可用来解轴对称问题，为了便于套用既有结果，把变量r, z改记为x, y，并且象14.3.1节中一样，把问题(14.4.1)稍稍推广为

$$\begin{cases} J(u)=\iint\limits_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}2\pi x\,dx\,dy \\ \qquad +\int_{\Omega'}\left[\frac{1}{2}\gamma_i u^2-qu\right]2\pi x\,ds+\sum\limits_{\Omega''}[(-pu)2\pi x]=\text{极小} \\ \Omega_0:\ u=\bar{u} \end{cases} \tag{14.4.5}$$

在rz平面内采用三角剖分和线性插值，同于§14.2，不同之点只是现在要计算基函数及其导数的乘积以x为权的积分，因此要用§14.2的表14.2, 14.4。应该记住，这里的三角面元所代表的实际上是三维空间里具有三角剖面的环体。相应的单元分析如下

1. 面元$C=(A_1, A_2, A_3)$

$$J_C=\frac{1}{2}\iint\limits_{C}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]-fu\right\}2\pi x\,dx\,dy\sim\sum_{i,j=1}^{3}a_{ij}^{(C)}u_iu_j-\sum_{i=1}^{3}b_i^{(C)}u_i$$

$$a_{ij}^{(C)}=2\pi x_0(\beta\eta_i\eta_j+\beta\xi_i\xi_j)/2D_0+2\pi(3x_0+x_i+x_j)\gamma D_0(1+\delta_{ij})/120$$

$$b_i^{(C)}=2\pi(3x_0+x_i)fD_0/24,\quad x_0=\frac{1}{3}(x_1+x_2+x_3)$$

2. 线元 $B=(A_1,\ A_2)$

$$J_B=\int_B\left[\frac{1}{2}\eta u^2-qu\right]2\pi x ds\sim\sum_{i,j=1}^{2}a_{ij}^{(B)}u_iu_j-\sum_{i=1}^{2}b_i^{(B)}u_i$$

$$a_{ij}^{(B)}=2\pi(x_0+x_i\delta_{ij})\eta L/6$$

$$b_{ij}^{(B)}=2\pi(2x_0+x_i)qL/6,\ x_0=\frac{1}{2}(x_1+x_2)$$

3. 点元 $A=(A_1)$

$$J_A=[(-pu)2\pi x]_A=\frac{1}{2}a_{11}^{(A)}u_1u_1-b_1^{(A)}u_1$$

$$a_{11}^{(A)}=0$$

$$b_1^{(A)}=2\pi x_0p,\quad x_0=x_1$$

此处 x_1 就是点 A_1 的 x 坐标。

总体合成方法与14.3.2节同。

注意在参考平面(r,z)中的面、线、点通过旋转一般地生成三维空间中的回转体、面、线，即上升一维。但也有例外，如原问题中还含有集中于 z 轴上的线源或点源项，即在能量积分(14.4.5)中再增加线项 $\int(-pu)dz$ 及点项 $\Sigma(-gu)$，当转化到参考平面中去时，这些项保持不变，应该加到能量积分(14.4.5)中去。注意它们不含有因子 $2\pi r$，这是因为在 Γ_s 上的点和线经旋转后保持不变，并不上升一维。对于(14.4.5)中增添的这些项的处理全同于14.3.1节。

对于轴对称问题还有简化的处理方案，即单元系数 a_{ij}, b_i 照用§14.3的公式，但普遍乘以因子 $2\pi x_0$, x_0 为每个单元的顶点 x_i(即 r_i)的算术平均值。

14.4.2 本征值问题

连续介质振动系统的自振频率和振型问题归结于椭圆方程的本征值问题。它和边值问题相仿，也有等价的变分原理，对此有限元法是同样适用的。

以弹性膜为例，取 $u(x,y)$ 为平衡态时弹性位移，β 为膜内张力(给定的系数)，f 为载荷分布，则膜的平衡方程就表为(14.1.16)中的第一式。(14.1.16)中的第二、三两式则表达了边界上以及内部交界上的平衡条件，q 表示边界上的线状载荷分布，η 表示边界弹性支承的弹性系数。在动态时，命 $w=w(x,y,t)$ 表示弹性位移，则膜体的运动方程是

$$\Omega:\ \rho\frac{\partial^2 w}{\partial t^2}-\left(\frac{\partial}{\partial x}\beta\frac{\partial w}{\partial x}+\frac{\partial}{\partial y}\beta\frac{\partial w}{\partial y}\right)=f$$

比平衡方程多了一个惯性力项 $\rho\frac{\partial^2 w}{\partial t^2}$, $\rho=\rho(x,y)$ 为单位面积的质量，相应的边界条件仍旧，即

$$L:\ \left(\beta\frac{\partial w}{\partial \nu}\right)^-=\left(\beta\frac{\partial w}{\partial \nu}\right)^+$$

$$\Gamma_0':\ \beta\frac{\partial w}{\partial \nu}+\eta w=q$$

$$\Gamma_0:\ w=\bar{w}$$

在作自由振动时，所有的载荷及强加条件都为0，因此方程和边值条件都成为齐次的，即

$$\Omega: \ \rho \frac{\partial^2 w}{\partial t^2} - \left(\frac{\partial}{\partial x}\beta \frac{\partial w}{\partial x} + \frac{\partial}{\partial y}\beta \frac{\partial w}{\partial y}\right) = 0$$

$$L: \ \left(\beta \frac{\partial w}{\partial \nu}\right)^- = \left(\beta \frac{\partial w}{\partial \nu}\right)^+$$

$$\Gamma_0': \ \beta \frac{\partial w}{\partial \nu} + \gamma_i w = 0$$

$$\Gamma_0: \ u = 0$$

在形成驻波的时候，解 $w(x, y, t)$ 可以表为

$$w(x, y, t) = e^{i\omega t} u(x, y)$$

ω 为自振频率，$u(x, y)$ 为相应的振型，以此代入上式即得关于 ω 及 u 的方程(命 $\lambda = \omega^2$)。

$$\begin{cases} \Omega: \ -\left(\dfrac{\partial}{\partial x}\beta \dfrac{\partial u}{\partial x} + \dfrac{\partial}{\partial y}\beta \dfrac{\partial u}{\partial y}\right) = \lambda \rho u \\ L: \ \left(\beta \dfrac{\partial u}{\partial \nu}\right)^- = \left(\beta \dfrac{\partial u}{\partial \nu}\right)^+ \\ \Gamma_0': \ \beta \dfrac{\partial u}{\partial \nu} + \gamma_i u = 0 \\ \Gamma_0: \ u = 0 \end{cases} \tag{14.4.6}$$

这样一组齐次方程显然有解 $u \equiv 0$，但是这种零解在物理上是不感兴趣的。关键在于仅当参数 λ 取某些特定值(叫做本征值)时，这组方程才有非零解——叫做本征函数。所谓本征值问题就是要求定出本征值和相应的本征函数。在这里本征值给出膜的自振频率 $\omega = \sqrt{\lambda}$，相应的本征函数给出振型。

命

$$D(u) = \iint_\Omega \frac{1}{2}\left[\beta \left(\frac{\partial u}{\partial x}\right)^2 + \beta \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy + \int_{\Gamma_0'} \frac{1}{2} \gamma_i u^2\, ds \tag{14.4.7}$$

$$E(u) = \iint_\Omega \frac{1}{2} \rho u^2\, dx\, dy \tag{14.4.8}$$

这是两个二次齐次泛函。可以证明，本征值问题(14.4.6)等价于下列“商”泛函的变分问题[1]

$$\begin{cases} J(u) \equiv \dfrac{D(u)}{E(u)} = \text{临界值} = \lambda & (14.4.9) \\ \Gamma_0: \ u = 0 & (14.4.10) \end{cases}$$

这就是说，在一切满足边界条件(14.4.10)并且不恒为 0 的函数类中使 J 达到临界的函数 u 就是本征函数，相应的值 $J(u)$ 即临界值就是本征值。

按照有限元法，在三角剖分下，通过单元分析和总体合成，二次泛函 D, E 可以分别离散化成为两个二次齐次函数

$$D(u) \sim D(u_1, \cdots, u_{N_0}) = \frac{1}{2} \sum_{i,j=1}^{N_0} a_{ij} u_i u_j \tag{14.4.11}$$

$$E(u) \sim E(u_1, \cdots, u_{N_0}) = \frac{1}{2} \sum_{i,j=1}^{N} c_{ij} u_i u_j \tag{14.4.12}$$

如果采用三角剖分和线性插值则有单元分析公式(14.3.1 节)。

1. 面元 $C=(A_1, A_2, A_3)$

$$D_C(u)=\iint_C \frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy\sim\frac{1}{2}\sum_{i,j=1}^{3}a_{ij}^{(C)}u_iu_j$$

$$E_C(u)=\iint_C \frac{1}{2}\rho u^2\,dx\,dy\sim\frac{1}{2}\sum_{i,j=1}^{3}c_{ij}^{(C)}u_iu_j$$

$$a_{ij}^{(C)}=\iint_C\left[\beta\frac{\partial\lambda_i}{\partial x}\frac{\partial\lambda_j}{\partial x}+\beta\frac{\partial\lambda_i}{\partial y}\frac{\partial\lambda_j}{\partial y}\right]dx\,dy=(\beta\eta_i\eta_j+\beta\xi_i\xi_j)/2D_0=a_{ji}^{(C)}$$

$$c_{ij}^{(C)}=\iint_C \rho\lambda_i\lambda_j\,dx\,dy=\rho D_0(1+\delta_{ij})/24$$

2. 线元 $B=(A_1, A_2)$

$$D_B(u)=\int_B\frac{1}{2}\eta u^2\,ds\sim\frac{1}{2}\sum_{i,j=1}^{2}a_{ij}^{(B)}u_iu_j$$

$$E_B(u)=0=\frac{1}{2}\sum_{i,j=1}^{2}c_{ij}^{(B)}u_iu_j$$

$$a_{ij}^{(B)}=\int_B\eta\lambda_i\lambda_j ds=\rho D_0(1+\delta_{ij})/24=a_{ji}^{(B)}$$

$$c_{ij}^{(B)}=0$$

仿照 14.3.2 节的原则进行累加就得到两个二次型(14.4.11～12)然后按照 14.3.3 节，设强加的零边界条件 (14.4.10) 作用于节点 $A_{h_1}, A_{h_2}, \cdots, A_{h_M}$，即

$$u_{h_i}=0,\quad i=1, 2, \cdots, M \tag{14.4.13}$$

从 a_{ij}, c_{ij} 两阵各自删去第 $h_1, \cdots, h_M$ 行和相应的列，同时将变数 $u_1, \cdots, u_{N_0}$ 删去相应的分量，设对余下的各元素按原顺序重新编号，就得到两个新的二次型(为了方便仍沿用原来的记号)

$$D(u_1, \cdots, u_N)=\sum_{i,j=1}^{N}a_{ij}u_iu_j,\quad a_{ij}=a_{ji} \tag{14.4.14}$$

$$E(u_1, \cdots, u_N)=\sum_{i,j=1}^{N}c_{ij}u_iu_j,\quad c_{ij}=c_{ji} \tag{14.4.15}$$

这里 $N=N_0-M_1$ 阵 $\boldsymbol{A}=[a_{ij}]$ 通常为正定或半正定，$\boldsymbol{C}=[c_{ij}]$ 为正定。于是变分问题(14.4.9～10)就离散化成为多元商函数的临界值问题

$$J(u_1, \cdots, u_N)\equiv\frac{D(u_1, \cdots, u_N)}{E(u_1, \cdots, u_N)}=\text{临界值}=\lambda \tag{14.4.16}$$

所谓 $(u_1, \cdots, u_N)\neq 0$ 使 J 达到临界是指 $(u_1, \cdots, u_N)$ 满足临界方程

$$\frac{\partial}{\partial u_i}J(u_1, \cdots, u_N)=0,\quad i=1, \cdots, N \tag{14.4.17}$$

相应的 J 值叫做临界值，记为 λ。由于

$$\frac{\partial}{\partial u_i}J=\frac{\partial}{\partial u_i}\left(\frac{D}{E}\right)=\frac{1}{E^2}\left(\frac{\partial E}{\partial u_i}D-\frac{\partial D}{\partial u_i}E\right)$$

$$=\frac{1}{E}\left(\frac{\partial E}{\partial u_i}J-\frac{\partial D}{\partial u_i}\right)=\frac{1}{E}\left(\frac{\partial E}{\partial u_i}\lambda-\frac{\partial D}{\partial u_i}\right)$$

并且 $(u_1, \cdots, u_N)\neq 0$，故(14.4.17)相应于

$$\frac{\partial E}{\partial u_i}\lambda-\frac{\partial D}{\partial u_i}=0,\quad i=1,\ \cdots,\ N$$

由于

$$\frac{\partial}{\partial u_i}D(u_1,\ \cdots,\ u_N)=\sum_{j=1}^{N}a_{ij}u_j,\quad \frac{\partial}{\partial u_i}E(u_1,\ \cdots,\ u_N)=\sum_{j=1}^{N}c_{ij}u_j$$

故(14.4.17)可以表为

$$\sum_{j=1}^{N}a_{ij}u_j=\lambda\sum_{j=1}^{N}c_{ij}u_j,\quad i=1,\ \cdots,\ N \tag{14.4.18}$$

用矩阵记号则为

$$\boldsymbol{Au}=\lambda\boldsymbol{Cu} \tag{14.4.19}$$

这组齐次方程仅当 λ 取一些特定值(叫做本征值)时才有非零解——叫做本征向量。这样，微分方程本征值问题(14.4.6)最终离散化为代数本征值问题(14.4.19)。对于后者有标准的数值解法，见第十章。

14.4.3 平面弹性问题

平面弹性问题在物理上有两类，即平面应变问题和平面应力问题，两者有统一的数学形式。

设在 x, y 方向的平面位移分布为 $u(x, y)$, $v(x, y)$，由此派生应变张量 ε_{xx}, ε_{xy}, ε_{yx}, ε_{yy} 和应力张量 σ_{xx}, σ_{xy}, σ_{yx}, σ_{yy}。

$$\varepsilon_{xx}=\frac{\partial u}{\partial x},\quad \varepsilon_{xy}=\varepsilon_{yx}=\frac{1}{2}\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right),\quad \varepsilon_{yy}=\frac{\partial v}{\partial y} \tag{14.4.20}$$

应变和应力张量之间有下列关系即虎克定律

$$\begin{cases}\sigma_{xx}=\alpha\varepsilon_{xx}+(\alpha-2\beta)\varepsilon_{yy}=\alpha\dfrac{\partial u}{\partial x}+(\alpha-2\beta)\dfrac{\partial v}{\partial y}\\[2mm] \sigma_{yy}=(\alpha-2\beta)\varepsilon_{xx}+\alpha\varepsilon_{yy}=(\alpha-2\beta)\dfrac{\partial u}{\partial x}+\alpha\dfrac{\partial v}{\partial y}\\[2mm] \sigma_{xy}=\sigma_{yx}=2\beta\varepsilon_{xy}=\beta\left(\dfrac{\partial u}{\partial y}+\dfrac{\partial v}{\partial x}\right)\end{cases} \tag{14.4.21}$$

α, β 为介质系数，可以依赖于 x, y，甚至可以有间断。在平面应变问题中

$$\alpha=\frac{E(1-\nu)}{(1-2\nu)(1+\nu)},\quad \beta=\frac{E}{2(1+\nu)} \tag{14.4.22}$$

E 为介质的杨氏模量，ν 为波瓦松比。在平面应力问题亦即薄板的纵向(板内)变形问题中

$$\alpha=\frac{Eh}{1-\nu^2},\quad \beta=\frac{Eh}{2(1+\nu)} \tag{14.4.23}$$

$h=h(x, y)$ 为板的厚度。在两种情况下应力张量的物理解释是有所不同的，在此不去深究。

在平面内任取弧长单元 ds(如图 14.10 所示)，规定其法向余弦为 ν_x, ν_y，切向余弦为 $\tau_x=-\nu_y$, $\tau_y=\nu_x$，于是位于 ds 正法向一侧通过 ds 作用于负法向一侧的弹性力(以单位长度计)在 x, y 方向的投影为

$$\nu_x\sigma_{xx}+\nu_y\sigma_{xy},\quad \nu_x\sigma_{yx}+\nu_y\sigma_{yy} \tag{14.4.24}$$

在法向及切向的投影则为

$$\sigma_{xx}\nu_x^2+2\sigma_{xy}\nu_x\nu_y+\sigma_{yy}\nu_y^2,\quad \sigma_{xx}\nu_x\tau_x+\sigma_{xy}(\nu_x\tau_y+\nu_y\tau_x)+\sigma_{yy}\nu_y\tau_y \tag{14.4.25}$$



图 14.10　　　　图 14.11

在弹性体所占的区域 Ω 内任取一个子域 D(图 14.11)，设在 x, y 方向的平面载荷分布为 f_x, f_y，于是有

$$\begin{aligned}&\oint_{\partial D}(\nu_x\sigma_{xx}+\nu_y\sigma_{xy})ds=\iint_D f_x\,dx\,dy\\&\oint_{\partial D}(\nu_x\sigma_{yx}+\nu_y\sigma_{yy})ds=\iint_D f_y\,dx\,dy\end{aligned}\tag{14.4.26}$$

这就是积分形式的平衡方程，ν_x, ν_y 表示 ∂D 上的外法向余弦。利用高斯积分公式就可以导此微分形式的平衡方程

$$\Omega:\ \begin{cases}-\left(\dfrac{\partial\sigma_{xx}}{\partial x}+\dfrac{\partial\sigma_{xy}}{\partial y}\right)=f_x\\[2mm]-\left(\dfrac{\partial\sigma_{yx}}{\partial x}+\dfrac{\partial\sigma_{yy}}{\partial y}\right)=f_y\end{cases}\tag{14.4.27}$$

这里应力分量 σ_{xx}, σ_{xy}, σ_{yx}, σ_{yy} 用 $\dfrac{\partial u}{\partial x}$, $\dfrac{\partial u}{\partial y}$, $\dfrac{\partial v}{\partial x}$, $\dfrac{\partial v}{\partial y}$ 的表达式(14.4.21)代入，即得到两个未知函数 u, v 的二阶椭圆型方程组。为了定解，应在边界 $\partial\Omega$ 给定两个边界条件。一般地可以把 $\partial\Omega$ 分解为三个互补的部分

$$\partial\Omega=\Gamma_0+\Gamma_1+\Gamma_2$$

在 Γ_0 上，位移全固定，取已知的分布

$$u=\bar{u},\quad v=\bar{v}\tag{14.4.28}$$

在 Γ_1 上，位移半固定。常见的条件是固定法向位移

$$u_\nu=\nu_x u+\nu_y v=\bar{u}_\nu\tag{14.4.29}$$

这时还要补充一个切向应力的边界条件，其一般形式有如

$$\sigma_{xx}\nu_x\tau_x+\sigma_{xy}(\nu_x\tau_y+\nu_y\tau_x)+\sigma_{yy}\nu_y\tau_y=-\eta(\tau_x u+\tau_y v)+q_\tau\tag{14.4.30}$$

恒约定 ν_x, ν_y 为外法向余弦，$\tau_x=-\nu_y$, $\tau_y=\nu_x$ 为切向余弦，(ν,τ) 构成一个局部的右手坐标系。右端 $-\eta(\tau_x u+\tau_y v)=-\eta u_\tau$ 表示切向的弹性反力，弹性系数 $\eta\geqslant 0$, q_τ 为线状切向载荷。

在 Γ_2 上，位移全自由。这时需要补充两个应力边界条件，其一般形式有如

$$\begin{aligned}\nu_x\sigma_{xx}+\nu_y\sigma_{xy}&=-(\eta_{xx}u+\eta_{xy}v)+q_x\\\nu_x\sigma_{yx}+\nu_y\sigma_{yy}&=-(\eta_{yx}u+\eta_{yy}v)+q_y\end{aligned}\tag{14.4.31}$$

弹性系数 η_{xx}, η_{xy}, η_{yx}, η_{yy} 形成一个对称半正定矩阵，给 x 及 y 方向的弹性反力，q_x, q_y 为两个方向的线状载荷。

当介质系数 α, β 有间断时，在其间断线 L 上通常假定两侧的位移连续

$$u^-=u^+,\quad v^-=v^+\tag{14.4.32}$$

这时尚应满足两个交界条件即应力平衡方程

$$
\begin{aligned}
(\nu_x\sigma_{xx}+\nu_y\sigma_{xy})^- &= (\nu_x\sigma_{xx}+\nu_y\sigma_{xy})^+ \\
(\nu_x\sigma_{yx}+\nu_y\sigma_{yy})^- &= (\nu_x\sigma_{yx}+\nu_y\sigma_{yy})^+
\end{aligned}
\tag{14.4.33}
$$

有时在 Ω 内部有缝隙。有一种是接触的缝隙 L_1，在其两侧法向位移连续

$$
(\nu_x u+\nu_y v)^- = (\nu_x u+\nu_y v)^+ \tag{14.4.34}
$$

而切向自由，可有滑移。这时应满足一个交界条件，即法向应力平衡

$$
(\sigma_{xx}\nu_x^2+2\sigma_{xy}\nu_x\nu_y+\sigma_{yy}\nu_y^2)^- = (\sigma_{xx}\nu_x^2+2\sigma_{xy}\nu_x\nu_y+\sigma_{yy}\nu_y^2)^+ \tag{14.4.35}
$$

还可以有脱离接触的缝隙 L_2，即两侧位移完全自由。这时应分别满足

$$
\begin{aligned}
(\nu_x\sigma_{xx}+\nu_y\sigma_{xy})^- &= 0, \quad (\nu_x\sigma_{xx}+\nu_y\sigma_{xy})^+ = 0 \\
(\nu_x\sigma_{yx}+\nu_y\sigma_{yy})^- &= 0, \quad (\nu_x\sigma_{yx}+\nu_y\sigma_{yy})^+ = 0
\end{aligned}
\tag{14.4.36}
$$

即无应力状态。

以上设缝隙 L_1, L_2 都相当窄，几何上可以视为相重合。

可以证明，以上的平衡方程连同其全部边界条件等价于下列变分问题即最小势能原理：

$$
\begin{cases}
J(u,\ v) = \displaystyle\iint_\Omega \left\{\frac{1}{2}[\alpha(\varepsilon_{xx}+\varepsilon_{yy})^2+4\beta(\varepsilon_{xy}^2-\varepsilon_{xx}\varepsilon_{yy})]-(f_x u+f_y v)\right\}dx\,dy \\
\qquad +\displaystyle\int_{\Gamma_1}\left\{\frac{1}{2}\eta_\tau(\tau_x^2u^2+2\tau_x\tau_y uv+\tau_y^2v^2)-q_\tau(\tau_x u+\tau_y v)\right\}ds \\
\qquad +\displaystyle\int_{\Gamma_2}\left\{\frac{1}{2}(\eta_{xx}u^2+2\eta_{xy}uv+\eta_{yy}v^2)-(q_x u+q_y v)\right\}ds = \text{极小} \\
\Gamma_0:\quad u=\bar{u},\ v=\bar{v} \\
\Gamma_1:\quad \nu_x u+\nu_y v=\bar{u}_\nu \\
L_1:\quad (\nu_x u+\nu_y v)^- = (\nu_x u+\nu_y v)^+
\end{cases}
\tag{14.4.37}
$$

注意所有关于应力的边界条件和交界条件都是自然边界条件，在变分问题中可以不列，因此情况大大简化。此外，在介质间断线 L 上约定位移取单值，因此位移连续条件(14.4.30)保证满足，故不作为强加条件列出。反之，在缝隙 L_1, L_2 约定位移取双值，在 L_1 上有一个约束(14.4.31)，作为强加条件列出，在 L_2 上则无约束。此外，在 Γ_1 上的积分可以统一为 Γ_2 上的形式，例如 $\eta_{xx}=\eta_\tau\tau_x^2$, $\eta_{xy}=\eta_\tau\tau_x\tau_y$, $\eta_{yy}=\eta_\tau\tau_y^2$, $q_x=q_\tau\tau_x$, $q_y=q_\tau\tau_y$。

对称性处理

当问题具有一定的对称性时，可以把定解区域简缩。应该指出这里有两个变量 u, v, 是同一个位移向量的两个分量。所谓位移对称性是指位移向量的对称性。例如说位移左右对称(对称于直线 $x=0$)是指位移向量对于镜射变换 $x\to -x$, $y\to y$ 为不变，即分量 u 反对称而分量 v 对称(图 14.12)

$$
\begin{aligned}
u(-x,\ y) &= -u(x,\ y) \\
v(-x,\ y) &= v(x,\ y)
\end{aligned}
$$

图 14.12

见图。显然可见，只有当定解区域，介质系数以及 y 方向的载荷和位移边界条件为左右对称以及 x 方向的载荷和位移边界条件为左右反对称时才能保证位移向量场的左右对称性。这时定解区间可以简缩一半。注意对称轴 $x=0$ 本来不是边界而在简缩后成为边界。在其上由 u, v 的正反对称性得到

$$x=0:\ u=0,\ \frac{\partial v}{\partial x}=0$$

其中第一个 $u=0$ 是强加条件，在简缩后需要增补为强加增补，第二个 $\frac{\partial v}{\partial x}=0$ 是自然边界条件，不必增补。一般说来，在作对称性简缩时，在对称轴上应增补一个法向位移为零的强加条件。

离散化

仍旧采用三角剖分，$u(x, y)$, $v(x, y)$ 在各单元上分别采用线性插值。为了方便，将各点元 $A_1, A_2, \cdots, A_{N_0}$ 的位移

$$u_1,\ v_1,\ u_2,\ v_2,\ \cdots,\ u_{N_0},\ v_{N_0}$$

统一记为

$$w_1,\ w_2,\ w_3,\ w_4,\ \cdots,\ w_{2N_0-1},\ w_{2N_0}$$

即

$$u_k=w_{2k-1},\quad v_k=w_{2k} \tag{14.4.38}$$

缝隙的处理同于14.3.3节所述，即缝上的点元和线元都一分为二，分别给予编号和位移，仅其几何坐标相同。

1. 面元分析 $C=(A_1, A_2, A_3)$

$$J_C=\iint_C\left\{\frac{1}{2}[\alpha(\varepsilon_{xx}+\varepsilon_{yy})^2+4\beta(\varepsilon_{xy}^2-\varepsilon_{xx}\varepsilon_{yy})]-(f_x u+f_y v)\right\}dx\,dy$$

把积分号下的二次项表为对称形式

$$\begin{aligned}\alpha(\varepsilon_{xx}+\varepsilon_{yy})^2+4\beta(\varepsilon_{xy}^2-\varepsilon_{xx}\varepsilon_{yy})&=\alpha\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)^2+\beta\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)^2-4\beta\frac{\partial u}{\partial x}\frac{\partial v}{\partial y}\\&=\alpha\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2+(\alpha-2\beta)\frac{\partial u}{\partial x}\frac{\partial v}{\partial y}+\beta\frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\\&\quad+\beta\frac{\partial v}{\partial x}\frac{\partial u}{\partial y}+(\alpha-2\beta)\frac{\partial v}{\partial y}\frac{\partial u}{\partial x}+\beta\left(\frac{\partial v}{\partial x}\right)^2+\alpha\left(\frac{\partial v}{\partial y}\right)^2\end{aligned}$$

用线性插值

$$u\sim\sum u_i\lambda_i,\quad \frac{\partial u}{\partial x}\sim\sum u_i\frac{\partial\lambda_i}{\partial x},\quad \frac{\partial u}{\partial y}\sim\sum u_i\frac{\partial\lambda_i}{\partial y}$$

$$v\sim\sum v_i\lambda_i,\quad \frac{\partial v}{\partial x}\sim\sum v_i\frac{\partial\lambda_i}{\partial x},\quad \frac{\partial v}{\partial y}\sim\sum v_i\frac{\partial\lambda_i}{\partial y}$$

α, β, f_x, f_y 均取常数值。于是

$$J_C\sim\frac{1}{2}\sum_{i,j=1}^{3}[a_{ij}^{(1)}u_iu_j+a_{ij}^{(2)}u_iv_j+a_{ij}^{(3)}v_iu_j+a_{ij}^{(4)}v_iv_j]-\sum_{i=1}^{3}[b_i^{(1)}u_i+b_i^{(2)}v_i]$$

$$a_{ij}^{(1)}=a\eta_i\eta_j+b\xi_i\xi_j,\quad a_{ij}^{(2)}=c\eta_i\eta_j+d\xi_i\xi_j$$

$$a_{ij}^{(3)}=d\eta_i\eta_j+c\xi_i\xi_j,\quad a_{ij}^{(4)}=b\eta_i\eta_j+a\xi_i\xi_j$$

$$b_i^{(1)}=f_xD_0/6,\qquad b_i^{(2)}=f_yD_0/6$$

$$a=\alpha/2D_0,\ b=\beta/2D_0,\ c=-(\alpha-2\beta)/2D_0,\ d=-b=-\beta/2D_0$$

若 $C=(A_{n_1}, A_{n_2}, A_{n_3})$，顶点标号为 n_i 的位移 u, v 已统一记为 w_{2n_i-1}, w_{2n_i}, $i=1, 2, 3$, 因此合成累加公式是

$$a_{ij}^{(1)}+a_{2n_i-1,2n_j-1}\Rightarrow a_{2n_i-1,2n_j-1}$$
$$a_{ij}^{(2)}+a_{2n_i-1,2n_j}\Rightarrow a_{2n_i-1,2n_j}$$
$$a_{ij}^{(3)}+a_{2n_i,2n_j-1}\Rightarrow a_{2n_i,2n_j-1}$$
$$a_{ij}^{(4)}+a_{2n_i,2n_j}\Rightarrow a_{2n_i,2n_j}$$
$$b_i^{(1)}+b_{2n_i-1}\Rightarrow b_{2n_i-1}$$
$$b_i^{(2)}+b_{2n_i}\Rightarrow b_{2n_i}$$
$$i,\ j=1,\ 2,\ 3$$

2. 线元分析 $B=(A_1,\ A_2)$

可以统一考虑为 Γ_2 上的形式即

$$J_B=\int_B\left\{\frac{1}{2}(\eta_{xx}u^2+2\eta_{xy}uv+\eta_{yy}v^2)-(q_xu+q_yv)\right\}ds$$

用线性插值

$$u\sim\sum u_i\lambda_i,\quad v\sim\sum v_i\lambda_i$$

η_{xx}, η_{xy}, η_{y}, q_x, q_y 均取常数值，于是

$$J_B=\frac{1}{2}\sum_{i,j=1}^{2}[a_{ij}^{(1)}u_iu_j+a_{ij}^{(2)}u_iv_j+a_{ij}^{(3)}v_iu_j+a_{ij}^{(4)}v_iv_j]-\sum_{i=1}^{2}[b_i^{(1)}u_i+b_i^{(2)}v_i]$$

$$a_{ij}^{(1)}=\eta_{xx}L(1+\delta_{ij})/6,\quad a_{ij}^{(2)}=\eta_{xy}L(1+\delta_{ij})/6$$
$$a_{ij}^{(3)}=\eta_{xy}L(1+\delta_{ij})/6,\quad a_{ij}^{(4)}=\eta_{yy}L(1+\delta_{ij})/6$$
$$b_i^{(1)}=q_xL/2,\qquad b_i^{(2)}=q_yL/2$$

若 $B=(A_{n_1},\ A_{n_2})$，即顶点标号为 n_1，n_2 时，合成累加公式与面元情况相同，不予赘述，但 $i,\ j=1,\ 2$。

3. 点元分析 $A=(A_1)$

有些问题的能量积分可能含有点项如

$$J_A=\left[\frac{1}{2}(\mu_{xx}u^2+2\mu_{xy}uv+\mu_{yy}v^2)-(p_xu+p_yv)\right]_A$$

μ_{xx}, μ_{xy}, μ_{yy} 为点弹性支承系数，p_x, p_y 为点载荷。这已经是离散的形式。命点 $A=A_1$ 的位移为 u_1, v_1,

$$J_A=\frac{1}{2}(a_{11}^{(1)}u_1v_1+a_{11}^{(2)}u_1v_1+a_{11}^{(3)}v_1u_1+a_{11}^{(4)}v_1v_1)-(b_1^{(1)}u_1+b_1^{(2)}v_1)$$

$$a_{11}^{(1)}=\mu_{xx},\quad a_{11}^{(2)}=a_{11}^{(3)}=\mu_{xy},\quad a_{11}^{(4)}=\mu_{yy}$$
$$b_1^{(1)}=p_x,\quad b_1^{(2)}=p_y$$

若 $A=(A_{n_1})$ 即顶点标号为 n_1 则合成累加方式也同面、线元相同，但 $i,\ j=1$。

关于强加条件的处理，这里要比 14.3.3 节复杂。

对于 Γ_0 上的点元，设其标号为 k，则(14.4.28)表为

$$w_{2k-1}=\bar{u}_k,\quad w_{2k}=\bar{v}_k \tag{14.4.39}$$

对于 Γ_1 上的点元，设其标号为 k，则(14.4.29)表为

$$\nu_{x,k}w_{2k-1}+\nu_{y,k}w_{2k}=\bar{u}_{\nu,k} \tag{14.4.40}$$

在 L_1 上的每点有“正”“负”两个点元，设其编号为 k, l，则(14.4.34)表为

$$\nu_{x,k}w_{2k-1}+\nu_{y,k}w_{2k}-\nu_{x,l}w_{2l-1}-\nu_{y,l}w_{2l}=0 \tag{14.4.41}$$

(14.4.39～41)一起构成了强加条件方程组，如 14.3.3 节中所述的一般形式(14.3.14)。

因此，可以用逐个分析强加条件点的方法来逐步形成有关的条件系数阵 $\boldsymbol{C}$, $\boldsymbol{d}$, 阵 $\boldsymbol{C}$ 当然也是稀疏的。按照拉格朗日乘子法，最终定解的系数阵(14.3.16)是

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}^T \\ \boldsymbol{C} & 0 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix}$$

考虑一种重要的特殊情况，即在边界上不受任何强加的条件。这时(14.4.37)成为无条件变分问题

$$J(u,\ v)=\iint_{\Omega}\left\{\frac{1}{2}[\alpha(\varepsilon_{xx}+\varepsilon_{yy})^2+4\beta(\varepsilon_{xy}^2-\varepsilon_{xx}\varepsilon_{yy})]-(f_x u+f_y v)\right\}dx\,dy$$

$$-\oint_{\partial\Omega}(q_x u+q_y v)\,ds=\text{极小} \tag{14.4.42}$$

这是退化半正定的情况。对应于(14.4.22)的齐次问题——即命 $f_x\equiv f_y\equiv 0$, $q_x\equiv q_y\equiv 0$

$$J_0(u,\ v)=\iint_{\Omega}\frac{1}{2}[\alpha(\varepsilon_{xx}+\varepsilon_{yy})^2+4\beta(\varepsilon_{xy}^2-\varepsilon_{xx}\varepsilon_{yy})]dx\,dy=\text{极小} \tag{14.4.43}$$

有三个非零基本解，

$$\begin{cases} u\equiv 1 \\ v\equiv 0 \end{cases}\quad \begin{cases} u\equiv 0 \\ v\equiv 1 \end{cases}\quad \begin{cases} u=y \\ v=-x \end{cases} \tag{14.4.44}$$

任意非零解可以表为这三个基本解的线性组合，

$$\left.\begin{aligned} u&=a+cy \\ v&=b-cx \end{aligned}\right\} \tag{14.4.45}$$

这表示不受约束不受载荷的弹性体的平衡解可以是也只能是刚性平移加旋转。

在非齐次的情况，问题(14.4.42)有解的充要条件即协调条件是

$$\iint_{\Omega} f_x\,dx\,dy+\int_{\partial\Omega} q_x\,ds=0 \tag{14.4.46}$$

$$\iint_{\Omega} f_y\,dx\,dy+\int_{\partial\Omega} q_y\,ds=0 \tag{14.4.47}$$

$$\iint_{\Omega}(yf_x-xf_y)\,dx\,dy+\int_{\partial\Omega}(yq_x-xq_y)\,ds=0 \tag{14.4.48}$$

这表示只有当外载荷的合力和合力矩为零时，不受约束的弹性体才能达成平衡。当此协调条件被满足时，任意两个解可以相差一个刚性运动即(14.4.45)。

以上变分问题(14.4.42)或(14.4.43)自然是在一切具有一定光滑性使积分(14.4.42)有意义的位移函数类 S 中定解的。在离散化后则是在 S 的子类(即一切片状线性的位移函数类)中定解，问题(14.4.42)或(14.4.43)分别变为

$$\text{在 } S \text{ 中定}(u,\ v)\text{使得}\quad J(u,\ v)=\text{极小} \tag{14.4.49}$$

$$\text{在 } S \text{ 中定}(u,\ v)\text{使得}\quad J_0(u,\ v)=\text{极小} \tag{14.4.50}$$

它们又分别等价于非齐次或齐次的线代数方程组

$$\begin{cases} \displaystyle\sum_{j=1}^{N_0} a_{ij}^{(1)}u_j+\sum_{j=1}^{N_0} a_{ij}^{(2)}v_j=b_i^{(1)} \\ \displaystyle\sum_{j=1}^{N_0} a_{ij}^{(3)}u_j+\sum_{j=1}^{N_0} a_{ij}^{(4)}u_j=b_i^{(2)} \end{cases} \tag{14.4.51}$$

$$\begin{cases} \sum_{j=1}^{N_0} a_{ij}^{(1)} u_j + \sum_{j=1}^{N_0} a_{ij}^{(2)} v_j = 0 \\ \sum_{j=1}^{N_0} a_{ij}^{(3)} u_j + \sum_{j=1}^{N_0} a_{ij}^{(4)} v_j = 0 \end{cases} \tag{14.4.52}$$

注意齐次变分问题(14.4.43)在 S 内的通解(14.4.45)是线性的，因此也属于其子类 S'，因此离散化后的齐次问题(14.4.50)即(14.4.52)的通解同样是(14.4.45)，同时也说明了离散问题(14.4.49)和(14.4.42)一样也是退化半正定，而且具有相同的"退化度"3——相当于三个基本解(14.4.44)。按照线代数的理论，非齐次问题(14.4.49)即(14.4.51)有解的充要条件为右项向量($\boldsymbol{b}^{(1)}$, $\boldsymbol{b}^{(2)}$)与齐次问题的基本解向量(即将(14.4.44)离散化)相正交，因此得三个协调条件

$$\sum_{i=1}^{N_0} b_i^{(1)} = 0 \tag{14.4.53}$$

$$\sum_{i=1}^{N_0} b_i^{(2)} = 0 \tag{14.4.54}$$

$$\sum_{i=1}^{N_0} (y_i b_i^{(1)} - x_i b_i^{(2)}) = 0 \tag{14.4.55}$$

类似于(14.3.25)，可以证明，如果单元分析中涉及 f_x, f_y, q_x, q_y 的计算是准确进行的话，则有

$$\sum_{i=1}^{N_0} b_i^{(1)} = \iint_\Omega f_x \, dx\, dy + \int_{\partial\Omega} q_x \, ds \tag{14.4.56}$$

$$\sum_{i=1}^{N_0} b_i^{(2)} = \iint_\Omega f_y \, dx\, dy + \int_{\partial\Omega} q_y \, ds \tag{14.4.57}$$

$$\sum_{i=1}^{N_0} (y_i b_i^{(1)} - x_i b_i^{(2)}) = \iint_\Omega (y f_x - x f_y) \, dx\, dy + \int_{\partial\Omega} (y q_x - x q_y) \, ds \tag{14.4.58}$$

因此原始的协调条件(14.4.46～48)自动保证离散的协调条件(14.4.53～55)，从而保证离散方程组(14.4.52)有解而任意两个解相差一个刚性运动(14.4.45)。这里再一次显示了有限元法在"特性保持"方面的优点。

在实践上，由于对 f_x, f_y, q_x, q_y 作了近似处理而条件(14.4.53～55)可能不严格成立，

$$\sum_{i=1}^{N_0} b_i^{(1)} = \varepsilon_1$$

$$\sum_{i=1}^{N_0} b_i^{(2)} = \varepsilon_2$$

$$\sum_{i=1}^{N_0} (y_i b_i^{(1)} - x_i b_i^{(2)}) = \varepsilon_3$$

这时可对 $b_i^{(1)}$, $b_i^{(2)}$ 加以调整即

$$b_i^{(1)} - a - c y_i \Rightarrow b_i^{(1)}$$

$$b_i^{(2)} - b + c x_i \Rightarrow b_i^{(2)}, \quad i = 1, \cdots, N_0$$

这里常数 a, b, c 是方程组

$$\begin{cases} a N_0 + c \sum_{i=1}^{N_0} y_i = \varepsilon_1 \\ b N_0 - c \sum_{i=1}^{N_0} x_i = \varepsilon_2 \\ a \sum_{i=1}^{N_0} y_i - b \sum_{i=1}^{N_0} x_i + c \sum_{i=1}^{N_0} (y_i^2 - x_i^2) = \varepsilon_3 \end{cases}$$

的解，而新的 b_i 满足(14.4.53～55)。有关退化平面弹性问题的代数解法可以参考[4]。

14.4.4 二次插值的应用

三角元的二次插值法(14.2.5 节)与线性插值法一样，对于二阶椭圆型问题包括边值问题和本征值问题是普遍适用的。下面仅以 §14.3 的问题(14.3.1)(稍加推广)为例来说明

$$\begin{cases} J(u)=\displaystyle\iint_{\Omega}\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2+\gamma u^2\right]-fu\right\}dx\,dy \\ \qquad +\displaystyle\int_{\Omega'}\left[\frac{1}{2}\eta u^2-qu\right]ds+\sum_{\Omega''}[-pu]=\text{极小} \\ \Omega_0:\quad u=\bar{u} \end{cases}$$

进行离散化时，首先注意插值节点除了全部点元外还包括全部线元的中点。需要对全部插值节点进行编号，也就是要对全部点元和全部线元进行统一的编号，相应未知数就是

$$u_1,\ u_2,\ \cdots,\ u_N,\quad N=N_0+N_1$$

参照 14.3.1 和 14.2.5 节可得单元分析和合成公式如下：

1．面元 $C=(A_1,\ A_2,\ A_3)$

$$u\sim\sum_{i=1}^{6}u_i\varphi_i,\quad \frac{\partial u}{\partial x}\sim\sum_{i=1}^{6}u_i\frac{\partial\varphi_i}{\partial x},\quad \frac{\partial u}{\partial y}\sim\sum_{i=1}^{6}u_i\frac{\partial\varphi_i}{\partial y}$$

$$J_C(u)=\iint_C\left\{\frac{1}{2}\left[\beta\left(\frac{\partial u}{\partial x}\right)^2+\beta\left(\frac{\partial u}{\partial y}\right)^2+\gamma u^2\right]-fu\right\}dx\,dy\sim\frac{1}{2}\sum_{i,j=1}^{6}a_{ij}^{(C)}u_iu_j-\sum_{i=1}^{6}b_i^{(C)}u_i$$

$$a_{ij}^{(C)}=\iint_C\left[\beta\frac{\partial\varphi_i}{\partial x}\frac{\partial\varphi_j}{\partial x}+\beta\frac{\partial\varphi_i}{\partial y}\frac{\partial\varphi_j}{\partial y}+\gamma\varphi_i\varphi_j\right]dx\,dy=a_{ji}^{(C)}$$

$$b_i^{(C)}=\iint_C f\varphi_i\,dx\,dy$$

关于基函数的积分问题见后。

若 C 的三顶点的统一编号为 $n_1,\ n_2,\ n_3$ 而相应的对边中点的统一编号为 $n_4,\ n_5,\ n_6$ 时，则在合成时应按下式累加

$$a_{ij}^{(C)}+a_{n_in_j}\Rightarrow a_{n_in_j}$$

$$b_i^{(C)}+b_{n_i}\Rightarrow b_{n_i},\quad i,\ j=1,\ \cdots,\ 6$$

2．线元 $B=(A_1,\ A_2)$

$$u\sim\sum_{i=1}^{6}u_i\varphi_i,\quad \frac{\partial u}{\partial s}\sim\sum_{i=1}^{6}u_i\frac{\partial\varphi_i}{\partial s}$$

$$J_B(u)=\int_B\left[\frac{1}{2}\eta u^2-qu\right]ds\sim\frac{1}{2}\sum_{i,j=1}^{3}a_{ij}^{(B)}u_iu_j-\sum_{i=1}^{3}b_i^{(B)}u_i$$

$$a_{ij}^{(B)}=\int_B\eta\varphi_i\varphi_j\,ds=a_{ji}^{(B)}$$

$$b_i^{(B)}=\int_B q\varphi_i\,ds$$

若 B 的两顶点统一编号为 $n_1,\ n_2$，而其中点的统一编号为 n_3 时，则合成时的累加公式为

$$a_{ij}^{(B)}+a_{n_in_j}\Rightarrow a_{n_in_j}$$

$$b_i^{(B)}+b_{n_i}\Rightarrow b_{n_i},\quad i,\ j=1,\ 2,\ 3$$

3．点元 $A=(A_1)$

$$u=u_1$$

$$J_A=\left[\frac{1}{2}\mu u^2-pu\right]_A=\frac{1}{2}a_{11}^{(A)}u_1u_1-b_1^{(A)}u_1$$

$$a_{11}^{(A)}=\mu$$

$$b_1^{(A)}=p$$

若点元 A 的统一编号为 n_1，则合成时的累加公式为

$$a_{11}^{(A)}+a_{n_1n_1}\Rightarrow a_{n_1n_1}$$

$$b_1^{(A)}+b_{n_1}\Rightarrow b_{n_1}$$

在以上单元系数的积分表达式事实上是一般的，在不同的插值方法中，只是基函数选取的不同。关于积分的计算，正如在 14.2.5 节末段所指出，可以采取数值积分的方法，它有通用的优点，便于把不同的插值方法统一在一个程序里。这里基函数及其导数都有重心坐标的表达式(14.2.15～16)，(14.2.18～19)，因此可用 14.2.4 节表 14.5～6 所列的适当精度的数值积分公式。即使对于线性插值也可以采用数值积分，而不用 14.3.1 节中所列的单元系数的明显公式。

对于三角剖分线性插值，每个面元有三个对应于顶点的未知数，总体方程的未知数总数为 N_0，即点元个数。在二次插值，每个面元上未知数从三个增至六个，但这并不意味着总体未知数比线性情况增至二倍，事实上，这时未知数总数为 N_0+N_1。根据近似比例 $N_0:N_2:N_1\approx1:2:3$(见 14.2.1 节)可知 $N_0+N_1\approx N_0+3N_0=4N_0$，增至四倍。因此，用线元中点参数插值一般是比较“费”的，但这只是相对于同一剖分而言。事实上，插值精度的提高意味着剖分有放粗的可能。实践表明，在达到同一合理精度要求的情况，用粗剖分二次插值比用细剖分一次插值的未知数总数往往要省得多，从而有可能大大压缩解题的规模。这一点对于本征值问题尤其重要。这是因为，当矩阵阶数较高而又要求定出多个本征值和本征向量时，计算方法上还有较大困难。

参 考 资 料

[1] 加藤敏夫，《变分法及其应用》，上海科学技术出版社，1961。

[2] 冯康，《基于变分原理的差分格式》，应用数学与计算数学，2:4(1965)，238～262 页。

[3] 齐基威茨一邱，《结构和连续力学中的有限单元体法》，国防工业出版社，1973。

[4] 黄鸿慈，王荩贤，崔俊芝，赵静芳，林宗楷，《按位移解平面弹性问题的差分方法》，应用数学与计算数学，3:1(1966)，54～60 页。

附录 算法语言 BCY 简介

BCY 是汉语拼音 Bianyi Chengxu Yuyan(编译程序语言)的缩写。BCY 语言在很多方面都与 ALGOL 语言相似，但是删去了 ALGOL 中的一些不太常用的成分(如递归，固有量等)，不再把量区分为整数型和实数型；此外，增加了一些成分，满足了符号加工和数位运算的需要，并使数组与数组之间的传送比较方便。这里将通俗而扼要地介绍算法语言 BCY㊀，其目的是帮助读者看懂本书前面一些章节中所附的程序。本附录是以参考资料[1]为依据编写的。

§1 概　述

本节是算法语言 BCY 的一个缩影，从中可以看到用 BCY 描述计算问题的概貌。下面用一些简单的例子来加以说明：

例 1：依次计算

$$C=A+B$$

$$Z=\frac{X-Y}{X\cdot Y}+2.6025$$

$$r=\sqrt{x^2+y^2}$$

$$f=-1.5\sin x+\ln(1+|x|)$$

在 BCY 中，可用如下四个"计算语句"表示：

```
A+B⇒C;
(X-Y)/(X*Y)+2.6025⇒Z;
§SQRT(X↑2+Y↑2)⇒R;
-1.5*§SIN(X)+§LN(1+§ABS(X))⇒F
```

可以看出，计算结果一律写在右面，并用赋值号"⇒"代替等号，乘号改用"*"，幂次放在"↑"之后，此外，初等函数前面加上了专用符号"§"，其变量一律放在圆括号内。各个语句之间都用分号相隔。

例 2：计算函数值

$$Y=\begin{cases}1-X^2 & (当\ X<0)\\ 1+X^2 & (当\ X\geqslant 0)\end{cases}$$

要说明的是：X 是一系列复杂运算的结果，编制程序时无法预料其正负，应作两手准备。BCY 中的"条件语句"可以完成这一任务：

若 X<0 **则** 1-X↑2⇒Y **否** 1+X↑2⇒Y

这里的汉字 **若**、**则**、**否**是 BCY 特别约定的，也是一个基本符号。为突出这一点，印刷时一律用黑体，书写时在该字下划一横线，汉字符号约有 30 个。

㊀ 实际上是介绍 BCY-乙，因为本书的全部计算程序都采用 BCY-乙 编制。

例 3：计算

$$\begin{cases} X=\dfrac{1}{E-A}+B \\ Y=\dfrac{1}{E-A}-B \\ Z=\dfrac{E}{5.205} \end{cases}$$

要说明的是：E 和 A 都是前面计算的结果，有可能出现 $E-A=0$ 的情况，这时不必算 X，Y，仅需算 Z。在 BCY 中，可用"转语句"来实现：

若 E－A＝0 **则转** BB **否** 1/(E－A)⇒C;

C＋B⇒X; C－B⇒Y;

BB: E/5.205⇒Z

上面的"BB: "称为"标号"，"转 BB"称为"转语句"，其意义是终断正常的计算顺序，直接转去执行标号 BB 后面的语句。在上例中则跳过 C＋B⇒X 和 C－B⇒Y 两个计算语句。

例 4：设有 A, B 两组数，各由 50 个数组成。

$$A=(A_1,\ A_2,\ \cdots,\ A_{50}),\ B=(B_1,\ B_2,\ \cdots,\ B_{50})$$

要求计算

(1) $C_i=A_i+2B_i \quad (i=1,\ 2,\ \cdots,\ 50)$

(2) $S=A_1B_1+A_2B_2+\cdots+A_{50}B_{50}$

这时可用"循环语句"实现。对于(1)有

对于 I＝1 **到** 50 **步长** 1 **执行** A[I]＋2*B[I]⇒C[I]

应该注意的是：下标 I 永远写在方括号内，A[I] 称为"下标变量"。上例句指出 I 是"循环参数"，它从 1 开始，每次增加 1(步长)，直到 50 止；并且对于每一个 I 值，执行后面的那个语句(A[I]＋2*B[I]⇒C[I])都被执行一次。类似地，对于(2)的计算，可写作

0⇒S;

对于 I＝1 **到** 50 **步长** 1 **执行** A[I]*B[I]＋S⇒S

这里的 S 开始取值 0，其后对每个 I，把乘积 A[I]*B[I] 累加到 S 上，最后就是 50 个乘积之和。

容易明白，例 4 中的 A, B, C 都代表由 50 个数组成的一个数组(BCY 中称为场)，但是例 1 中的 A, B, C 却代表单个变量(BCY 中不再区分它们是实数还是整数，一律称为简变)。我们知道，在计算机作计算处理时，单个变量和一个数组有很大的差别。因此哪些量是单个变量，哪些量是数组，应在程序中分别"说明"。下面，我们用 BCY 对例 4 作较详细的描述：

```
始  简变 S;
    场 A, B, C[1:50];
    输十A, B;
    对于 I=1 到 50 步长 1 执行
        A[I]+2*B[I]⇒C[I];
    0⇒S;
    对于 I=1 到 50 步长 1 执行
        A[I]*B[I]+S⇒S;
```

印 S, C; 停 555

终

上面是一个完整的 BCY 程序，它以始开头，终结尾，并将“说明”放在前头。例中有两个说明，即“简变 S”和“场 A, B, C[1:50]”，它们指出本程序中的 S 是单个变量，而 A, B, C 则是各有 50 个数的数组。在 BCY 中，循环参数是不需要说明的。“输十A, B”称为“输入语句”，它把十进制数据 A, B(根据说明，它们各有 50 个数)输入计算机，等待进一步的运算处理。“印 S, C”称为“印刷语句”，它把结果 S 和 C(共 51 个数)以十进制形式打印出来，因此印也可以写为印十。停后的编号称为停机编号，不同的编号可以区分不同位置设置的停机。停机编号也可以略而不写。

§2 BCY 中的几种主要成分

在进一步介绍之前，首先指出：在程序中用以代表不同运算对象的名字(如 A, B, E, BB, I 等)，其最一般形式是以字母为开头的字母、数字的组合，如

$$\text{Beijing},\ x_{12},\ W_{4G},\ c\theta s,\ V_{\theta 1}$$

都是名字。BCY 中约定用 θ 代替字母 O，其目的是不与数 0 相混淆；此外，还约定字母不论大写、小写，也不论写在上半行或下半行，都看作是同一个符号，例如 a 与 A 相同，Mx 与 mx 相同，x_{12} 或 x^{12} 与 X12 相同。

2.1 条件语句

条件语句的一般形式是

若 E1～E2 则 S1 否 S2

其中“E1～E2”称为“条件”；“～”则是 <, ⩽, = 这三个关系符中的任何一个；S1 和 S2 都是一个任意的语句。当条件满足时，执行 S1(不执行 S2)，反之执行 S2(不执行 S1)。

下面是几点补充说明：

(1) 条件中的 E1, E2 可以是表达式。例如

2*N−1⩽M 或 0.001<§ABS(B[K]/4)

(2) S1 或 S2 中的任何一个可以是“空语句”(即不写任何符号)。例如，把 X 变为 X 的绝对值，可写作

若 X<0 则 −X⇒X 否(这里的否不能省去)

或

若 0⩽X 则否 −X⇒X

(3) 条件语句中的 S1 和 S2 还可以是条件语句。例如，要印出 A, B, C 三个数中最大的一个，则可写作：

若 A<B 则若 B<C 则 C⇒MAX 否 B⇒MAX
否若 A<C 则 C⇒MAX 否 A⇒MAX;

印 MAX

上面的第一个句子可理解为具有如下结构：

若 A<B 则 S1 否 S2

其中 S1 是： 若 B<C 则 C⇒MAX 否 B⇒MAX

S2 是：若 A<C 则 C⇒MAX 否 A⇒MAX

在计算过程中，将根据条件只执行 S1 或 S2。如果执行 S1，亦将根据条件只执行传送语句 C⇒MAX 或 B⇒MAX；如果执行 S2，则只执行 C⇒MAX 或 A⇒MAX，因此，不论执行了哪一个传送语句，都认为整个条件语句已被执行完毕，往下都应执行其后的印刷语句。

在实际计算中，条件语句可以层层套叠，组成很复杂的逻辑判断。

(4) 按定义，条件语句中则和否后面的 S1 和 S2 都是一个语句。但是，在某些条件下，可能需要同时执行多个语句，这时需引入复合语句和分程序的概念。

2.2 复合语句和分程序

复合语句的一般形式是：

始 S1；S2；…；SN 终

其中 SI(I=1, 2, …, N)是各种语句。

分程序的一般形式是：

始 D1；D2；…；DM；S1；S2；…；SN 终

其中 DI(I=1, 2, …, M)是各种说明，SJ(J=1, 2, …, N)是各种语句。根据定义，分程序与 §1 末尾所说的一个(完整)程序的形式完全相同。因此，一个分程序可以就是一个程序(在 BCY 中，一个复合语句也可以是一个程序)。需要强调的是：一个复合语句或一个分程序也可以看作是一个语句。因此条件语句中的则与否后面的那一个语句可以是一个复合语句或分程序。

例 1：有三个量 D0, D1, D2，若 D_0 为负，则需改变 D1 和 D2 的符号。这时可用 BCY 语句写作

若 D0<0 则始 −D1⇒D1；−D2⇒D2 终否

例中是用一个复合语句作为则后的语句。

例 2：有三个量 TA, TB, TC，若 TA 为负，则需交换 TB 与 TC 之值。这时可写作

```
若 0≤TA 则否
  始  简变 D;
      TB⇒D; TC⇒TB; D⇒TC
  终
```

例中是用一个分程序作为否后的语句。

一个复合语句或分程序可以看作是一个语句的概念极为重要，它们是构成更为复杂的语法现象的基础。其后，我们将不止一次地使用这个概念。

最后，我们指出，在一个分程序开头的说明，只在本分程序内有效，因此可以随意选用名字，而不管它是否与分程序外的名字相同。显然，该分程序与外面有联系的量不应在这个分程序内说明。

例：

```
始  简变 D, TA, TB, TC;
    输十 TA, TB, TC; (2*TB+TC)/TA⇒D;
    若 TA<0 则
      始  简变 D;
```

```
        TB⇒D; TC⇒TB; D⇒TC
    终否;
  印 D;
  ……
终
```

上面的程序是一个分程序形式，其中条件语中的则与否之间也是一个分程序。外分程序中把 D 和 TA 等量说明为简变，但内分程序中又把 D 重新作了说明。根据分程序内的说明只在本分程序内有效的原则，内分程序中的语句 TB⇒D 仅把 TB 值赋予内分程序说明的 D，而与外分程序说明的 D 毫不相干。因此外分程序中印出的 D 还是(2*TB+TC)/TA 的值。如果外分程序中对 D 没有说明，则条件语句后的印刷语句内不能出现 D。换句话说，在外分程序中的印刷语句不能直接印出在内分程序中说明的量。在内分程序中没有被重新说明的量(例中是 TB, TC)依然按外分程序的说明起作用。

2.3 场的说明和使用

在 §1 例 4 中，我们已经看到，场 A[1:50] 表示由 50 个数组成的一组数，下标变量 A[1], A[I+2] 分别表示这组数中的第 1 个和第 I+2 个数。下面再作进一步的介绍。

设平面上有 N 个点：$P_1(X_1, Y_1)$, $P_2(X_2, Y_2)$, $\cdots$, $P_N(X_N, Y_N)$。如果把这 N 个点的坐标排列成 X_1, Y_1, X_2, Y_2, $\cdots$, X_N, Y_N，并定义场 XYR[1:N, 1:2]，则下标变量 XYR[K, I] 表示第 K 个点的第 I 个坐标，例如 XYR[1, 2] 表示 Y_1，XYR[2, 1] 表示 X_2。如果把这 N 个点的坐标排列成 X_1, X_2, $\cdots$, X_N, Y_1, Y_2, $\cdots$, Y_N，并定义场 XYV[1:2, 1:N] 则下标变量 XYV[I, K] 表示第 K 个点的第 I 个坐标，例如 XYV[1, 2] 表示 X_2，XYV[2, 1] 表示 Y_1。

场内的元素参加运算，通常都是通过下标变量进行的。下标的个数应与场说明中的维数一致，并且下标的取值要在相应的界对范围内。例如有场 R[0:5, −1:N, 3:7]，则下标变量可写作 R[I, J, K]，其中 $0 \leqslant I \leqslant 5$, $-1 \leqslant J \leqslant N$, $3 \leqslant K \leqslant 7$。

下面三种特殊情况允许仅写场的名字：

(1) 输入与印刷语句(见 §1 末尾的程序)

(2) 表达式送场：一般形式是

E⇒A　(E 为表达式, A 为场)

其意义是把表达式 E 的值送给场 A 的每一个分量。例如有场 Y[1:2, 0:2]，则 0⇒Y 表示把 0 送给 Y 的 6 个分量中的每一个。

(3) 场送场：一般形式是

A⇒B　(A, B 都是场)

其意义是把场 A 的全部分量依次送到场 B 的相应分量中。一般要求 A, B 两个场的维数、大小相同，否则按下述原则处理：若 A 的分量个数比 B 少，则把 A 的全部分量依次送到 B 的前部；若 A 的分量比 B 的分量多，则把 A 的前部分量依次送 B，直到 B 被送满为止。

2.4 循环语句

循环语句中最常见的形式是

对于 I=A 到 B 步长 C 执行 S

其中 I 是循环参数，是一个名字；A 是循环的初值，B 是循环的终值，C 是每循环一次时 I 应增加的量。A, B, C 都可以是任意的表达式；S 称为循环体，它是一个任意语句(包括复合语句、分程序和循环语句)。执行上述循环语句与执行下述分程序等价：

```
始  简变 I, N, H;
    A⇒I; B⇒N; C⇒H;
L1: 若 0<(I-N)*SIGN(H) 则转 L2 否;
    S; I+H⇒I; 转 L1;
L2:
终
```

因为循环参数 I 已在等价的分程序内被说明了，因此不必(也不允许)另加说明。但根据分程序内说明的量仅在本分程序内有效的原则，循环参数仅在该循环语句内起作用，离开该循环就失去意义，这要特别加以注意。

由于循环体可以是一个任意的语句，因此可以出现很复杂的循环嵌套形式。

例：计算

$$C_i=\sum_{j=1}^{N} A_{ij}B_j \quad (i=1,\ 2,\ \cdots,\ M)$$

要指出的是，类似的计算共有 10 组，它们分别与参数 $K=-10,\ -20,\ \cdots,\ -100$ 对应。而且每组的 A_{ij} 与 B_j 都不同，甚至 $M,\ N$ 也不同。

下面是实现这个计算的一个完整程序。其中假设与各组 A_{ij}, B_j 相应的 $M,\ N$ 也作为原始数据一同输入，因此每组的数据共有 $M\cdot N+N+2$ 个，排列次序如下：

$$\underbrace{M,\ N,\ A_{ij},\ B_j}_{\text{第一组}(K=-10)},\ \underbrace{M,\ N,\ A_{ij},\ B_j}_{\text{第二组}(K=-20)},\ \cdots,\ \underbrace{M,\ N,\ A_{ij},\ B_j}_{\text{第十组}(K=-100)}$$

```
始  对于 K=-10 到 -100 步长 -10 执行
      始  简变 M, N;
          输十M, N;
            始场 A[1:M, 1:N], B[1:N], C[1:M];
              输十A, B;
              0⇒C;
              对于 I=1 到 M 步长 1 执行
                  对于 J=1 到 N 步长 1 执行
                      A[I, J]*B[J]+C[I]⇒C[I];
              印 K, C
            终
      终;
    停 33
终
```

下面是对这个程序的一些解说：

(1) 这个程序的开头没有说明，只有两个语句(循环语句和停语句)。这是复合语句构

成一个完整程序的例子。

(2) 第一个循环语句的循环体是个分程序。它由一个说明(简变 M, N)和两个语句组成,其中第一个语句是"输十M, N",第二个语句又是一个分程序(直至印 K, C 结束),这里使用分程序形式是必要的,因为场 A, B, C 的说明依赖于 M, N,因此必须放在输入 M, N 之后。如果场说明之前的"始"不写,将导致语句后直接出现说明,违反了关于说明必须集中放在一个分程序的前头的规定(见本附录 2.2)。

(3) 在内分程序中,有一个语句"0⇒C",它表示场 C 的 M 个分量的值都变为零。此外还有一个循环语句:

```
对于 J=1 到 N 步长 1 执行
    A[I, J]*B[J]+C[I]⇒C[I]
```

在语法分析时,这个语句被看作是另一个循环语句(对于 I=1 到 M 步长 1 执行…)的循环体。亦即

```
对于 I=1 到 M 步长 1 执行
    对于 J=1 到 N 步长 1 执行
        A[I, J]*B[J]+C[I]⇒C[I]
```

组成了两重循环,并且它们又被嵌套在最外面的循环(对于 K=-10 到 -100 步长 -10 执行…)之中。因此这个程序含有三重循环。

由于循环语句在执行时,总是先检查循环参数的当前值是否已超出终止值,当超出时该循环语句就算执行完毕,因此,下述语句在 P>10 时,实际上不作什么计算工作:

```
对于 L=P 到 10 步长 1.25 执行 S
```

同理,下述两个语句也是等价的:

```
对于 I1=0 到 25 步长 3 执行 S1;
对于 I1=0 到 24 步长 3 执行 S1
```

循环语句有更一般的形式:

```
对于 I=T 执行 S
```

其中 I 仍为循环参数,S 仍为一个任意语句组成的循环体,但 T 是一个"循环元表",它由一个或多个(其间用逗号分隔)"循环元"组成。循环元共有三种形式:

(1) E(E 为表达式)

(2) A 到 B 步长 C(A, B, C 为表达式)

(3) E 当 E1~E2(E, E1, E2 为表达式,~ 是 $<$, $\leqslant$, $=$ 这三个关系符中的任何一个)。

对于(2),最常见的是只有一个循环元,这种循环语句已在前面介绍过了。

根据(1),可以写出如下形式的语句:

```
对于 PH=1, 3, 10, 50 执行 0⇒A[PH]
```

这时将 0 送入 A[1], A[3], A[10], A[50]这四个下标变量中。

带条件的循环形式(3),其意义可用等价的语句来表示:执行循环语句

```
对于 I=E 当 E1~E2 执行 S
```

与执行下述分程序等价:

```
始  简变 I;
```

```
LL: E⇒I;
      若 E1~E2 则始 S; 转 LL 终否
终
```

要指出的是,在循环元表中,上述三种循环元可以混合出现。例如:

对于 J0=1 到 5 步长 1, 10 到 30 步长 5, 50, 100, 200 执行印 V[J0]

这时将印出 V[1], V[2], …, V[5], V[10], V[15], …, V[30], V[50], V[100], V[200]。

2.5 开关

为使程序有更大的灵活性,引进"开关"的概念。例如有一运算,它依赖于 M, N 两个参数,而这样的参数共有 3 组(譬如 1, 2; 4, 6 和 3, 7),至于在一次计算中应该使用哪组参数,则由前面计算所得的 X 值来决定(譬如 $X=1$ 则用第 1 组;$X=2$ 则用第 2 组,$X=3$ 则用第 3 组)。这时可用开关来实现:

```
始  简变 X;
    开关 SABC[AA, BB, CC];
    ......
    ...⇒X;
    转 SABC[X];
AA: 1⇒M; 2⇒N; 转 ABC;
BB: 4⇒M; 6⇒N; 转 ABC;
CC: 3⇒M; 7⇒N;
ABC: ......
终
```

上述程序中的"开关 SABC[AA, BB, CC]"称为开关说明,其中 SABC 是这个开关的名字,方括号内的 AA, BB, CC 都是标号名字。"转 SABC[X]"称为开关语句,它由 X 的值决定转向,若 $X=K$ 则转向开关说明中的第 K 个标号。例如 $X=2$,则转向标号 BB。

2.6 过程

不少的计算程序往往需要在不同的地方完成相同(或类似)的计算,这时可把相同部分的程序首先编出,并给它一个名字,其后需要这部分计算时,只需写出这个名字,这就是过程的概念。先编出的公共部分称为"过程说明",其后写此过程名字(即调用此过程)称为"过程语句"。例如,程序中需要多处计算复数除法:

$$Z_1+iZ_2=(X_1+iX_2)/(Y_1+iY_2)$$

这时可用过程实现:

```
始  简变 X1, X2, Y1, Y2, Z1, Z2;
    过程 DIV;
      始 Y1↑2+Y2↑2⇒Z2; (X1*Y1+X2*Y2)/Z2⇒Z1;
         (X2*Y1-X1*Y2)/Z2⇒Z2
      终;
```

```
    AA: …; DIV; …
        DIV; …; DIV; …
    终
```

上例中的

```
    过程 DIV; 始 Y1↑2+…⇒Z2 终
```

便是过程说明，它指出 DIV 是过程名字，其内容由紧接其后的那个语句(称为"过程体")给出。过程体可以是复合语句或分程序，也可以是一个普通语句。过程说明之后出现的过程名字 DIV 便是过程语句，执行它等价于执行过程体

```
    始 Y1↑2+…⇒Z2 终
```

过程说明也是说明中的一种，因此它应当与其它说明(如简变说明、场说明、开关说明等)一起放在分程序的开头。正式的计算工作是从说明部分之后的第一个语句开始的。上例就是从标号 AA 开始。AA 前的那个复合语句是属于过程说明的，如果其后设有过程语句 DIV 调用，它将不会被执行。

如果要对不同的复数作除法，譬如

$$C_1+iC_2=(A_1+iA_2)/(B_1+iB_2)$$
$$G_1+iG_2=(E_1+iE_2)/(F_1+iF_2)$$
$$P_1+iP_2=(C_1+iC_2)/(G_1+iG_2)$$

则上述形式的过程是不够方便的。为此引入有参数过程的概念。这时可写出如下程序

```
    始  过程 DIVI(X1, X2, Y1, Y2, Z1, Z2);
            值 X1, X2, Y1, Y2; 简变 Z1, Z2;
            始 Y1↑2+Y2↑2⇒Z2;
              (X1*Y1+X2*Y2)/Z2⇒Z1;
              (X2*Y1-X1*Y2)/Z2⇒Z2
            终;
        简变 A1, A2, B1, B2, C1, C2, E1, E2, F1, F2, G1, G2, P1, P2;
        ……
        DIVI(A1, A2, B1, B2, C1, C2);
        DIVI(E1, E2, F1, F2, G1, G2);
        DIVI(C1, C2, G1, G2, P1, P2);
        ……
    终
```

例中的

```
    过程 DIVI(X1, X2, Y1, Y2, Z1, Z2);
        值 X1, X2, Y1, Y2; 简变 Z1, Z2;
        始 Y1↑2+…⇒Z2 终
```

是有参数的过程说明。过程名字 DIVI 之后用一对圆括号指出该过程所依赖的参数 (X1, X2, Y1, Y2, Z1, Z2)，紧接其后的"值 X1, X2, Y1, Y2"和"简变 Z1, Z2"都称为参数的"种类部分"，用于指出这些参数属何种类("值"的意义将在稍后给出，这里不妨理解为与简变一样)，种类部分之后的那个复合语句仍然称为过程体。该过程体之后出现三个形式相近

的语句，其中第一句是

DIVI(A1, A2, B1, B2, C1, C2)

它表示调用**过程** DIVI，并且要用这里的参数 A1, A2, B1, B2, C1, C2 依次替换过程说明中的参数 X1, X2, Y1, Y2, Z1, Z2，其结果相当于执行如下的复合语句：

```
始 B1↑2+B2↑2⇒C2;
   (A1*B1+A2*B2)/C2⇒C1;
   (A2*B1-A1*B2)/C2⇒C2
终
```

即实现了复数除法

$$C_1+iC_2=(A_1+iA_2)/(B_1+iB_2)$$

类似地，其后的两个过程语句，也分别完成了复数除法，得到结果 G_1+iG_2 和 P_1+iP_2。

上例可以清楚地看出，过程说明中的参数 X1, X2, Y1, Y2, Z1, Z2 只是形式的，它们既可以是其后的 A1, A2, B1, B2, C1, C2, 也可以是 E1, E2, F1, F2, G1, G2 等，因此称为“形式参数”，而过程语句中的参数称为“实在参数”。

有参数的过程说明比较重要，特别是因为它把本过程所依赖的全部参数明显地展现出来，因此编制各种算法程序几乎都使用过程说明形式。

下面是对有参数过程的一些补充说明。

(1) 过程说明中的形式参数，其种类包括**值**、**简变**、**场**、**开关**和**过程**。形式参数仅在该过程说明中有效。过程语句中的实在参数，其种类应与形式参数一致。

(2) 形式参数为值时，实在参数可以是表达式(包括一个简变、下标变量或常数)。在 BCY 中，调用过程时，遇有**值**参数，先自动算出该表达式的值，并送入过程内部定义的一个简变中，其后，过程体内的有关运算仅与这个内部定义的简变打交道。值参数不得出现在赋值号⇒的右面。

(3) 形式参数为场时，种类部分不必指出其维数与大小。例如

```
始  过程 SUM(A, N, S); 场 A; 值 N; 简变 S;
      始 0⇒S;
         对于 I=1 到 N 步长 1 执行
             S+A[I]⇒S
      终;
    简变 D;
    场 XY[1:100];
    输十XY;
    SUM(XY, 100, D);
    印 D
终
```

尽管形式参数的种类部分中**场** A 的维数与大小不如场说明那样一目了然，但此过程对**场** A 的要求是明确的(即**场** A 是一维的，且至少有 N 个分量)。有关这类信息，一般都在该过程的使用介绍中一一指出。

(4) 开关可以作为过程参数。这主要是用于过程有非正常出口的情况。例如

```
过程 GV(CH, M, FAIL);
    场 CH; 值 M; 开关 FAIL;
    始……
      转 FAIL[1]; ……
      ……转 FAIL[2]; ……
    终
```

此过程在编写时已预料计算过程中可能有两种不正常情况出现(如迭代不收敛或矩阵 CB 的主元素为零等),但本过程不便作进一步的处理,需终止计算,交回调用它的程序处理。因此调用该过程时,应该有一个开关说明。譬如

开关 SW[F1, F2]

它与形式参数 FAIL 对应,并且标号 F1 和 F2 标出的语句,应能分别处理上述两种非正常情况。

由于 BCY-乙不允许直接使用标号作过程参数,因此本书中的一些程序常出现只有一个标号的开关作为过程参数。

(5) 一个过程可以作为另一个过程的参数,从而使过程具有更大的灵活性。

例: 用矩形法计算如下两个定积分的值:

$$S=\int_0^{10} e^{-x}\,dx \qquad (\text{步长 } 0.1)$$

$$S_1=\int_1^2 x^2/3.25\,dx \quad (\text{步长 } 0.02)$$

矩形法的一般公式是

$$S=\int_a^b f(x)\,dx=h\sum_{i=1}^{n} f(a+ih)$$

其中 n 为积分区间 (a, b) 的细分个数, $h=\dfrac{b-a}{n}$ 是步长。上述计算可用如下程序实现:

```
始 过程 INT(A, B, F, S, N);
    值 A, B, N; 过程 F; 简变 S;
  注 {本过程用矩形法计算定积分 S=∫_A^B F(x)dx, 其中 N 是区间(A, B)的细分个数, 过程 F(x, y) 用于计算被积
     函数,由使用者提供, x 是积分变量,函数值放在 y 中}
    始 简变 H, Y;
       0⇒S; (B−A)/N⇒H;
       对于 I=1 到 N 步长 1 执行
         始 F(A+I*H, Y); S+Y⇒S 终;
       H*S⇒S
    终; 注 {过程 INT 的说明完毕}
  过程 F0(X, Y); 值 X; 简变 Y;
     §EXP(−X)⇒Y; 注 {本过程仅一个语句}
  过程 F1(X, Y); 值 X; 简变 Y;
     X*X/3.25⇒Y; 注 {本过程仅一个语句}
  简变 S, S1; 注 {全部说明完}
```

```
    INT(0, 10, F0, S, 100);
    INT(1, 2, F1, S1, 50);
    印 S, S1
终
```

例中过程 INT 以过程 F 为形式参数，但种类部分中对过程 F 没有进一步的说明(也没有指出是否带参数，但在注中已给使用者作了明确的交待)。不过作为实在参数的过程 F0 和 F1 是有完整的说明的，并且它们已按注中提出的要求编制。

顺便指出，例中的名字 S，既是形式参数，也是实在参数，这是巧合，不能因此省去了实在参数 S 的说明，因为形式参数与实在参数完全不同，而且形式参数仅在过程体中有定义。根据本附录 2.2 最后例子指出的量的有效范围和同名量处理的原则，执行这里的第二个过程语句

INT(1, 2, F1, S1, 50)

时，尽管它调用的过程体中有形式参数 S，但决不影响第一个过程语句中已计算出来的实在参数 S 的值。因此打印出来的S还是 $\int_0^{10} e^{-x}\,dx$ 的值。

(6) BCY-乙中已有若干常用算法(如解线代数方程组等)的标准过程说明，使用者不必再作说明就可以调用。这种过程的名字之前都有一个专门的符号“§”，详见参考资料[1]。

参 考 资 料

[1] 109-乙机算法语言编译小组，《109-乙机算法语言及其编译程序使用说明》，科学出版社，1973。